Evaluation of the Sequencing Pipeline for the Oxford Nanopore MinION Long-read DNA Sequencer

Sampo Koivunen

Master's thesis

University of Helsinki

Faculty of Agriculture and Forestry

Helsinki Region Biotechnology Education Programme (HEBIOT)

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department	
Faculty of Agriculture and Forestry	Helsinki Region Biotechnology Education Programme HEBIOT	
Tekijä/Författare – Author		
Sampo Koivunen		
Työn nimi / Arbetets titel – Title		
Evaluation of the Sequencing Pipeline for the Oxford Nanopore MinION Long-read DNA Sequencer		
Oppiaine /Läroämne – Subject		
Biotechnology		
Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Master's Thesis	April 2019	84

Tiivistelmä/Referat – Abstract

The Oxford Nanopore MinION is a third generation sequencer utilizing nanopore sequencing technology. The nanopore sequencing method allows sequencing of either DNA or RNA strands as they pass through the membrane-embedded nanopores. By measuring the corresponding fluctuations in the ion flow passing through the nanopore the passing strands can be sequenced directly without additional second-hand reactions or measurements.

The MinION sequencing has very distinctly different characteristics compared to the market leaders of the sequencing field. The small form factor of the device further helps it to separate itself from the other alternatives. However, the technology has only been on the market for a very short time and thus very little golden standards regarding its capabilities or usage have been established.

This thesis describes our experiences testing the capabilities of the MinION sequencer both before its commercial release as a part of a special early access program, as well as our continued experiments with the device following its commercial launch. The main results of this study include successfully sequencing and aligning *E.coli* and human gDNA samples to their respective reference genomes. Using our sequencing and analysis pipeline specifically tuned to the MinION we were able to sequence the entire *E.coli* genome on a single MinION flow cell with an average depth of around 180.

Over the course of the thesis project the MinION sequencing protocol was evaluated and optimized in order to determine whether it has the potential to achieve our ultimate goal of reliably sequencing the previously inaccessible genomic regions of the human genome. The possibility of augmenting the sequencing protocol by adding the pre-sequencing target enrichment was also explored. Ultimately we were able to confirm that the MinION sequencer can be used to sequence long DNA fragments from a multitude of sample types. The majority of the produced reads could successfully be aligned against a reference genome. However, the limited yield and sequencing quality of a single experiment does limit the applicability of the method for more complicated genomic studies. These issues can be addressed with various techniques, chiefly target enrichment, but adapting such methods into the sequencing pipeline has its own challenges.

Avainsanat – Nyckelord – Keywords

NGS, third-generation sequencing, long-read sequencing, sequencing, nanopore, MinION, bioinformatics, data analysis, Oxford Nanopore Technologies

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department	
Maatalous-metsätieteellinen tiedekunta	Helsingin Seudun Biotekniikan Koulutusohjelma	
Tekijä/Författare – Author		
Sampo Koivunen		
Työn nimi / Arbetets titel – Title		
Nanopore MinION – sekvensointimenetelmä pitkille DNA-fragmenteille – menetelmän testaus ja arviointi		
Oppiaine /Läroämne – Subject		
Biotekniikka		
Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Maisterin tutkielma	Huhtikuu 2019	84

Tiivistelmä/Referat – Abstract

Oxford Nanopore MinION on uusi kolmannen sukupolven sekvensointilaite, jolla on monia erityispirteitä. MinION:n käyttämä nanopore -sekvensointimenetelmä perustuu havaittavissa olevien jännitevirtausten muutosten mittaamiseen DNA- tai RNA-juosteiden kulkeutuessa nanokokoisten huokosten eli porejen kautta membraanin läpi. Menetelmä mahdollistaa juosteiden suoran sekvensoimisen ilman välireaktioita.

Uniikista sekvensointitavastaan johtuen MinION -laitteen tyyppiominaisuudet ovat hyvin erilaiset kuin laajemmin käytössä olevilla sekvensointilaitteilla. Myös MinION -laitteen poikkeuksellisen pieni koko auttaa sitä entisestään erottumaan kilpailijoistaan. Nanopore-pohjainen sekvensointiteknologia ja MinION ovat kuitenkin olleet kaupallisesti saatavilla vasta lyhyen aikaa. Siksi menetelmä on vielä suurelta osin standardisoimaton ja sen sovellettavuutta tutkimuskäytössä ei pystytä vielä tarkasti arvioimaan.

Tässä pro gradu -työssä kuvataan MinION -sekvensoinnin käyttöönottoa sekä arvioidaan sen suorituskykyä. Työn käytännön tutkimus aloitettiin jo ennen laitteen kaupallista julkaisua markkinoille osana erillistä ennakkotestausohjelmaa nimeltä MinION Access Programme (MAP) ja se jatkui katkeamatta myös MinION:n kaupallisen lanseerauksen jälkeen. Tutkimuksessa sekvensoitiin sekä *E.coli*-kasvatuksesta että ihmisverestä eristettyjä gDNA-näytteitä. Tuloksena saadut sekvenssit oli pääosin mahdollista linjata referenssigenomeihin. Sekvensointi- ja analyysivaiheiden optimoinnin jälkeen yhdellä sirulla pystyttiin tuottamaan tarpeeksi sekvenssidataa kattamaan *E.coli*-genomi kokonaisuudessaan keskimääräisellä 180x lukusyvyydellä.

Tutkimuksessa arvioitiin MinION:n suorituskykyä tavoitteena arvioida, sopiiko menetelmä ihmisgenomin hankalasti sekvensoitavien alueiden luotettavaan tutkimiseen. Lisäksi testattiin mahdollisuutta täydentää sekvensointimenetelmää erillisellä protokollalla kohdennetun sekvensoinnin toteuttamiseksi. Tutkimuksen tulokset osoittavat, että MinION – menetelmää voidaan käyttää pitkien ja linjattavissa olevien sekvenssien tuottamiseen. Sirujen sekvensointikapasiteetti ja sekvenssien laatu kuitenkin rajoittavat menetelmän käytettävyyttä monimutkaisempien genomien tutkimuksessa. Kohdennusprotokollan ja muiden täydentävien menetelmien liittäminen osaksi sekvensointiprosessia voi auttaa näiden puutteiden ratkaisemisessa, mutta tällaisten laajennusprotokollien käyttöönotto saattaa olla haasteellista.

Avainsanat – Nyckelord – Keywords

NGS, kolmannen sukupolven sekvensointi, pitkäjuosteinen sekvensointi, sekvensointi, nanopore, MinION, bioinformatiikka, data-analyysi, Oxford Nanopore Technologies

Säilytyspaikka – Förvaringställe – Where deposited

Muita tietoja – Övriga uppgifter – Additional information

Contents

ABBREVIATIONS

ASIC	Application-Specific Integrated Circuit
CSC	IT Center for Science, a non-profit corporation offering computing resources for scientific use.
GUI	Graphical User Interface
HPC	High-Performance Computing
IDT	Integrated DNA Technologies (company)
kmer	A string of nucleotides of k length. Example: a 5-mer is a string of five nucleotides.
NEB	Nebulin (gene)
NGS	Next-Generation Sequencing
ONT	Oxford Nanopore Technologies (company)
RNN	Recurrent Neural Network
ROI	Region of Interest
TAITO-SHELL	TAITO computing environment of CSC
TRI	Nebulin triplicate region

1 INTRODUCTION

1.1 Preface

Ever since the first model of the biochemical structure of DNA was presented in 1953 the field of DNA sequencing has been a strongly contested topic in the field of genomics.[1] Accurate understanding of the genomic sequence of different species at the nucleotide level would permit researchers to study the principles of inheritance and genomic variance at an intimate level. This is something much coveted in the fields of hereditary and medical research. As is typical for any field of research combining wide-scale interest and immense amounts of scientific potential, the sequencing space has seen rapid and essentially uninterrupted advancement ever since its conception continuing to this day.

The rapid pace the methods of sequencing have been evolving has led to a widely accepted generational division between various sequencing technologies. This division is based on the large-scale advancements between the modern and legacy sequencing methods. While this rate of advancement is extremely impressive for a practical field of science, only in existence for some 50 years, it also serves as a great source of confusion and nomenclatural inconsistencies. The terminology has simply not been able to keep up with the pace of the multiple newly developed technologies and scientific breakthroughs. This has led to the current state of partially blurred generational definitions regarding the existing sequencing technologies.

Even the widely accepted convention of referring to the sequencing technologies capable of large-scale data production using the next-generation sequencing (NGS) label has become under some level of scrutiny. New devices, often superior in both their technology and sequencing capabilities, continue to be released on the market. Regardless of the specific, they are often somewhat forcefully included under the single NGS label, obfuscating the scientific discussion. This has caused the once descriptive generational distribution to lose its meaning over time. With many long-standing sequencing solutions and devices, the situation is even more muddled. Some have seen

enough personal advancement over time to arguably validate the notion of generational separation of their newer iterations from the predecessor models.

One of the more widely accepted clarifications to the sequencing technology generation issue is through renaming the NGS as the second sequencing generation instead. The first-generation would then refer to the legacy methods while the NGS label would be reserved to the most recent sequencing solutions, the ostensible "third generation". Sometimes the terms third and fourth generation are also used in similar manner, depending on how distinct the differences between sequencers are considered. Basic characteristics of some of the more common sequencing technologies are listed in Table 1. Listed is also one plausible generational segregation of the different technologies, adopted for the purpose of this thesis.

Table 1. Common sequencing technologies segmented by the generation. The basic properties of each method are listed in the table for comparison.

	Sequencing generation	Sequenceable sample types	Average run time	Max read length (bp)	Accuracy (%)
Sanger	First	DNA	20 min - 3 h	900	99,9
Illumina	Second	DNA	1-11 days	600	99,9
Pyrosequencing	Second	DNA	24 h	700	99,9
Ion torrent	Second	DNA	2 h	600	99,6
SOLiD	Second	DNA	1- 2 weeks	100	99,9
PacBio	Third	DNA	30 min - 20 h	>100 000	85–90
Nanopore	Third	DNA, RNA	30 min - 48 h	>2 Mbases	80-90

1.2 Motivations behind the study

The Oxford Nanopore MinION -sequencer is a newly established sequencing device based on the principle of nanopore sequencing, elements of which were first theorized as early as 1989.[2] The unique characteristics of the nanopore-based sequencing make it dramatically different from the other sequencing methods and paint the device with both a great deal of potential as well as its own unique shortcomings. These will be discussed in further detail in the *Literature review* section of this thesis.

The defining feature of the MinION nanopore sequencing is its radically different read profile compared to most of the established sequencing solutions. The methodological differences will be explained in detail in the other sections of this thesis. A typical nanopore read from the Oxford Nanopore MinION is somewhat inferior in quality to those produced by standard sequencing devices of the industry such as Illumina NextSeq. In sharp contrast to Illumina reads, MinION reads are not intrinsically or technologically limited in their length or the nucleotide composition of the target region. The sample preparation steps for a MinION sequencing run are fairly straightforward.

Furthermore, the limitations regarding the sequencing targets or sample types due to the sequencing method itself are rather minimal. In addition to these characteristics the small stature and high portability of the MinION device itself allow nanopore sequencing to immediately establish its own niche in the heavily contested sequencing spa

The potential of the nanopore sequencing is immense. For example, it is valued by research groups focusing on the research of genomic regions traditionally challenging to sequence with the current sequencing methods. One such region is the human nebulin gene (*NEB*) which is the most commonly mutated gene in the nemaline myopathy patients and an important research target. Nemaline myopathy is an inheritable neuromuscular disorder with varying degrees of severity.[3,4,5] The large size and characteristics of the *NEB* gene have made sequencing and research of some of these

variations extremely difficult to study. Among the most problematic are the variations of the *NEB* triplicate region. This large and highly repetitive region in the middle of the gene consists of three tandem repeats of eight consecutive exons (82–89, 90–97 and 98–105).[6] The *NEB* triplicate region has been observed to commonly contain copy number variations (CNV). The copy number gains of over one extra copy in this region have been suggested to be pathogenic.[7] Despite strong research interest towards the triplicate region the large size of roughly 32 kb and the high level of repetition have made solving the accurate nucleotide-level sequence unachievable using traditional sequencing methods.[7] Image 1 provides a more detailed depiction of the genomic structure of the *NEB* gene and the triplicate region.

While the MinION sequencer is an interesting proposition for sequencing the aforementioned *NEB* triplicate region and other similarly challenging regions of the human genome, its applicability for such a task must first be assessed thoroughly in practical laboratory environment through experimental testing. Aside from being the standard procedure whenever new technology is adopted for the purposes of scientific research, the MinION device had only just recently entered a MinION Access Programme (MAP) phase comparable to a limited-access beta release of a computer software during the planning stages of this study. As such, there also existed no ready-made sequencing pipelines or analysis solutions that could be used as basis for the applicability testing at the time. This thesis project was established in order to answer these challenges, become familiar with the device as well as the associated sequencing workflows and to reliably assess the capabilities of the MinION sequencer. The ultimate target of the project was to perform and analyze long-range DNA sequencing with the MinION device while simultaneously constructing a functional sequencing and analysis pipelines for future applications of the technology.

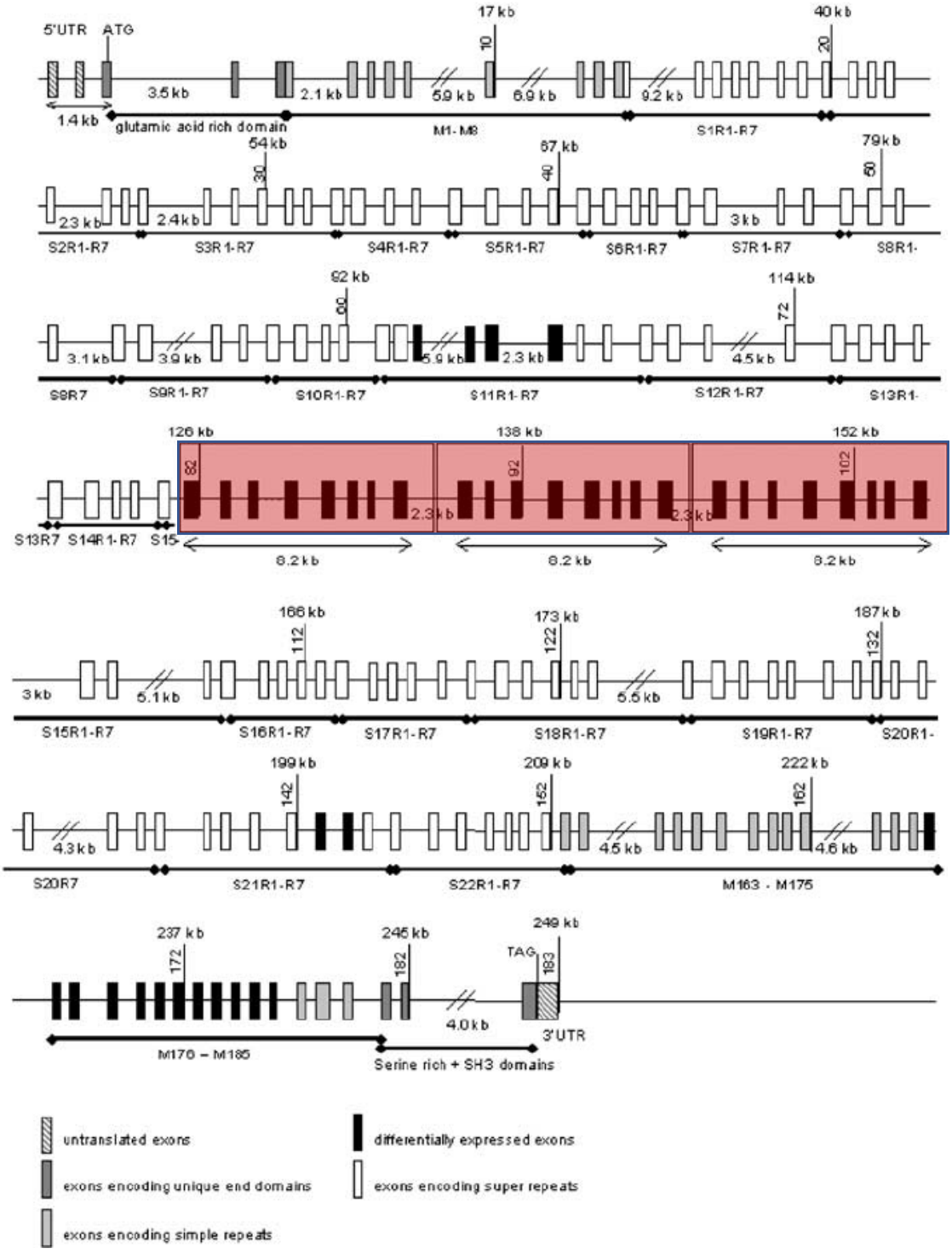

Image 1. The genomic structure of the human nebulin gene.[8]*The original image has been modified to better visualize the triplicate region of the gene with red highlighting.*

The progression route of this thesis project was strongly dependent on the progression of the development on the MinION device and the results produced by the sequencing experiments during the study. The more promising the results obtained with the MinION, the more complete and robust the resulting sequencing pipeline could become. Given an opportunity, more supportive elements such as target enrichment could also be included. Conversely, if the performance of the MinION were to be determined to be dissatisfactory, the further construction of the long-read sequencing pipeline would either be halted or the efforts would be redirected towards alternative sequencing options. As such, the end status of the final sequencing pipeline by the conclusion of the thesis project was deemed to be malleable by necessity, mainly determined by the success of the experiments themselves instead of a pre-established set of requirements.

Outside of the construction of the experimental method, a large amount of resources during the study were directed towards the research and exploration of applicable data analysis solutions for the data produced by the sequencing experiments. This was important to ensure the constructed sequencing pipeline was as independent of external research entities as possible. Furthermore, many of the traditional sequencing analysis methods were ill prepared and innately not well suited for processing the long and error-prone read profile of the nanopore reads. This was especially pronounced in the early stages of the project when both raw and basecalled sequence data produced by the MinION were stored using non-standard fast5 file format, essentially making them entirely unusably without the use of specialized analysis tools.

1.3 Aim of the study

Even though the apparent potential of the MinION device has been clear ever since ONT first presented nanopore sequenced DNA data in February 2012,[9] at the inception of this study it was immediately obvious that harnessing the technology for practical research was still a long time away. As such, the aim of the study was adjusted towards testing and evaluating the MinION device still under active development with the long-term goal of achieving an applicable sequencing pipeline for future studies. All laboratory and bioinformatics steps needed to produce usable sequencing data with the MinION device were to be included as part of this pipeline but should remain modular enough to allow their replacement later with alternative tools if necessary. The ultimate aim for this pipeline is the sequencing of aforementioned *NEB* triplicate region and other similar scientifically relevant genomic targets. For this thesis, producing sufficient sequencing data for performance analysis and completing the basic sequencing pipeline from sample preparation to result analysis was deemed sufficient. In addition to these requirements some preliminary work needed to advance the usability of the pipeline towards targeted sequencing was also performed and is included as part of this thesis to better assess the future of the technology.

2 LITERATURE REVIEW

2.1 The History of Nanopore Sequencing

The Oxford Nanopore MinION is the first commercially available sequencing tool utilizing nanopore technology for genomic strand sequencing. However, the idea of using nanoscale pores for nucleotide recognition itself is far from novel. The first mentions regarding the concept can be traced as far back as 1989 with the sequencing principle later described in more detail by Kasianowicz *et al.* in their 1996 published article regarding the subject.[10, 11]

The leading principle of the nanopore sequencing is the idea of driving the research sample, in the case of genomic sequencing a DNA strand, through a nanoscale pore in a controlled manner while simultaneously measuring the resulting ripple effects caused by this movement to the flow passing through the pore. The main driving forces of the sample movement are the electronic gradient over the pore-hosting membrane and an externally introduced voltage from a single side of the membrane. The introduction of voltage drives the gradient towards equilibrium and generates a stable flow observable as an electronic current through to the pore. As this electronic current is measured in real time with extremely sensitive local electrodes, the entering and passing of a DNA strand in the pore causes distinct fluctuations in the signal measurements. As these fluctuations are caused by the differing nucleotides passing through the pore over time, they can be identified and distinguished on a nucleotide level with a sensitive enough recognition algorithm and subsequently converted into sequence data. This principle is further illustrated in the images 2 and 3 and explained in more detail on the official ONT sources.[12]

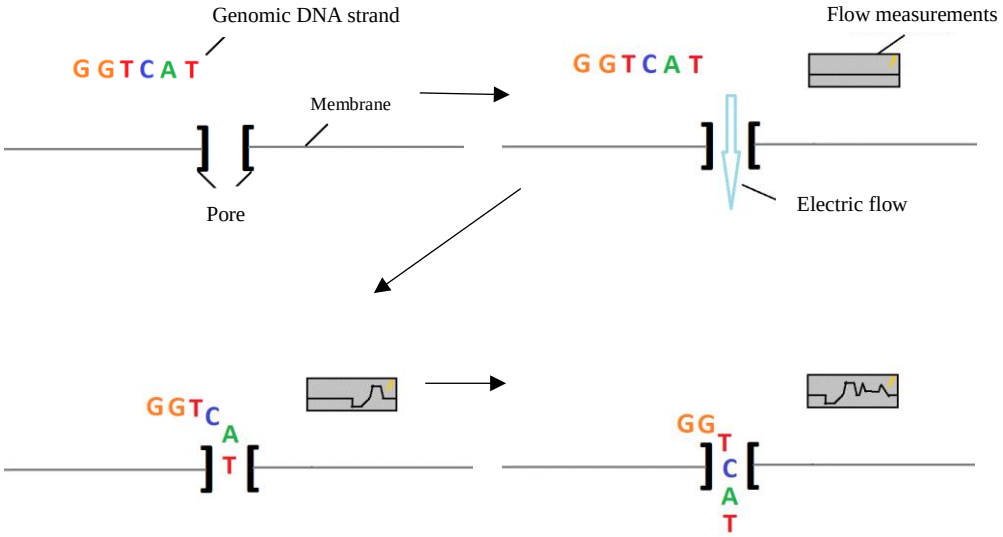

Image 2. A simplified visual representation of the nanopore sequencing principle.

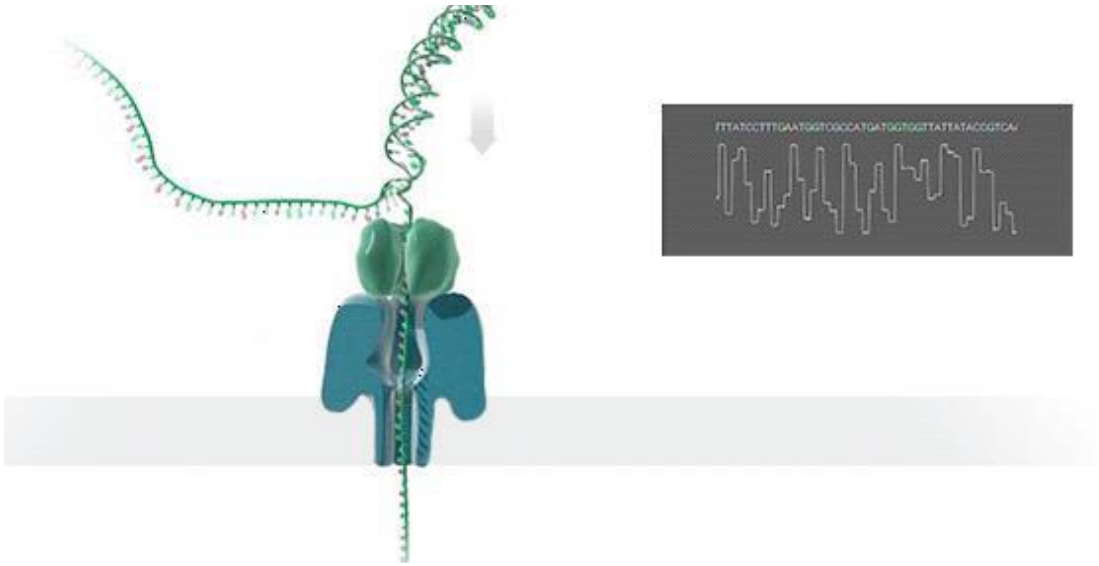

Image 3. A molecular model depicting the nanopore sequencing process in the MinION device.[12]

In fact, the major prohibitive factor for the commercialization of nanopore sequencing has not been the lack of theoretical understanding of the method, but the capability of producing and combining the specific set of fine structural and biochemical elements requirement for a functional sequencing device with capabilities to correctly produce sequencing data. Producing any sort of structure with nanoscale elements naturally comes with its innate challenges such as demanding structural work and issues regarding cost, reproducibility and quality assurance of the production. In the case of nanopore sequencing the structural challenges are further marred by the additional need for accurate electronic flow control and measurement process for all channels, each in the scale nanometers. The MinION sequencer is the first commercially available sequencing device to achieve this in practice.

2.2 Oxford Nanopore MinION sequencer

2.2.1 The MinION device

Although the Oxford Nanopore MinION Sequencer is recurrently treated as a single entity in this thesis, in truth the complete device configuration consists of two clearly distinct and individual mechanical components controlled using a separate computer software. The three separate entities: the MinION device, the replaceable flow cells and the MinKNOW sequencing software form together the fully operational MinION sequencer, working together in every sequencing run. However, all three parts are structurally entirely independent from each other. Updates and alterations can be introduced for each of these three components separately from the other two, as long as the high-level compatibility between the three is preserved. While such build structure is great from the developmental standpoint of the device and helps keeping down the cost of upgrades for end user, it may also confuse those unfamiliar with the technology.

The first and the most immediately recognizable part of the MinION sequencing setup is the MinION device as depicted in Image 4. For a sequencing device, it is very small, with dimensional measurements of around 105x35x25 mm and the weight of around 90 grams. The outward appearance of the device is fairly simplified, lacking any sort of screen or operational interface. The latest model has an openable lid connected by a single hinge to the short side of the device, hiding the insertion point for flow cells underneath it. The bottom part of the MinION case is dotted with small holes both facilitating the ventilation of the device and hiding underneath the status led lights activated during runtime. Underneath the metallic shell of the device are contained electronic components for control and measurement operations of the sequencing run, operated with the help of the MinKNOW software. The further details regarding the internal structure of the MinION remain unreleased by ONT and thus cannot be elaborated further in this thesis. Finally, on the opposite side of the lid hinge is the only external connection port of the device fitting a specialized USB 3.0 cable connector. It

is through this port that the MinION both receives all the power it requires and is connected to a computer during runtime.

Image 4. The MinION mk1B sequencer with the lid closed (left) and open with a flow cell attached (right).

Out of the three parts of the complete sequencer setup, the MinION device has been the most stable thus far. The device does not experience any particular wear-and-tear during the sequencing run and can be used repeatedly without any problems or need for component replacements. The original model, generally referred to as simply MinION1 in this thesis, was the only available version at the beginning of the study. It was replaced by a newer model MinION mk1 in May 2015, which was consequently superseded by MinION mk1B the following year, May 2016.[9] The observable changes of these MinION updates to the end user are much less pronounced compared to those resulting from the upgrades to the flow cells or MinKNOW software. The change from MinION mk1to the MinION mk1B was especially minor from optics viewpoint, as the outward appearance of both models is identical. The MinION1 was much more distinct in its outward appearance, sporting different designs for lid and flow cell insertion point. The detailed information regarding the internal changes between different versions is scarcely available but the functional differences between each model have been documented by the developer. These changes have been listed in the Table 2.

2.2.2 The MinION flow cell

Other than the MinION device, the second specialized physical component of the completed MinION sequencing setup is the replaceable flow cell. Unlike the MinION device, these flow cells are neither capable nor intended for repeated use. ONT does

offer a flow cell washing kit that can be used to remove previous sample post-sequencing and preserve the flow cell to be used again on another sequencing run. However, this washing is not entirely flawless and trace amounts of the previous sample is bound to be left within the flow cell even after the washing procedure, leading to a contamination risk in the later runs. Another factor limiting the performance of the washing procedure is the natural depletion of chemical balance and loss of pores occurring within the flow cell during every sequencing run. This has an expected negative impact on both the yield and read quality in consecutive runs on a single flow cell, which cannot be alleviated through the washing protocol.[12]

The structural construction of the flow cell combines biochemical and electronical components with nanoscale sized functional elements. Image 5 depicts the up- and downsides of a modern MinION flow cell with the different elements of the flow cell labelled. Out of all the parts comprising the flow cell, the ones with the most sophisticated structure and importance for the function are the sequencing chamber and Application-Specific Integrated Circuit (ASIC) operating the computational functions of the flow cell. These elements have bolded on the front and backside images respectively. The priming port is used to pipette the priming mix to the flow cell prior to the sequencing run. The SpotON sample port cover is protecting a SpotON port, through which the sample is loaded directly into the sequencing chamber. The other elements are mostly structural without notable functional properties and do not actively participate in the sequencing process.

Table 2. The changes of the Nanopore MinION sequencing system elements between iterations.

MinION		**Flow cell**		**Sequencing kit**		
Iteration	**Major changes**	**Iteration**	**Major changes**	**Iteration**	**Read type(s) supported**	**Changes to the sequencing protocol**
MiniION 1.0	Original version	FLO-MAP001	Original version	SQK-MAP002	1D and 2D	First version of the sequencing kit used in this study
MinION mk1	New design, simplified attachment of flow cells, support for sequencing enhancements	FLO-MAP002	General improvements	SQK-MAP005	1D and 2D	General improvements to the overall stability and performance of the library preparation protocol
MinION mk1B	Improvements in sequencing speed and quality	FLO-MAP003	General improvements	SQK-MAP006	1D and 2D	Doubled sequencing speed to the maximum of 70 b/s, new enzyme E6 replacing previous E5 enzyme and removal of HP motor protein
		FLO-MIN004	New R9 pore	SQK-MIN007	1D and 2D	New enzyme
		FLO-MIN106	New R9.4 pore, SpotON port	SQK-NSK007	1D and $1D^2$	New enzyme, increased sequencing speed to the maximum of 250 b/s and compatibility with R9 pore flow cells
				SQK-LSK108	1D and $1D^2$	General improvements to the stability and yield

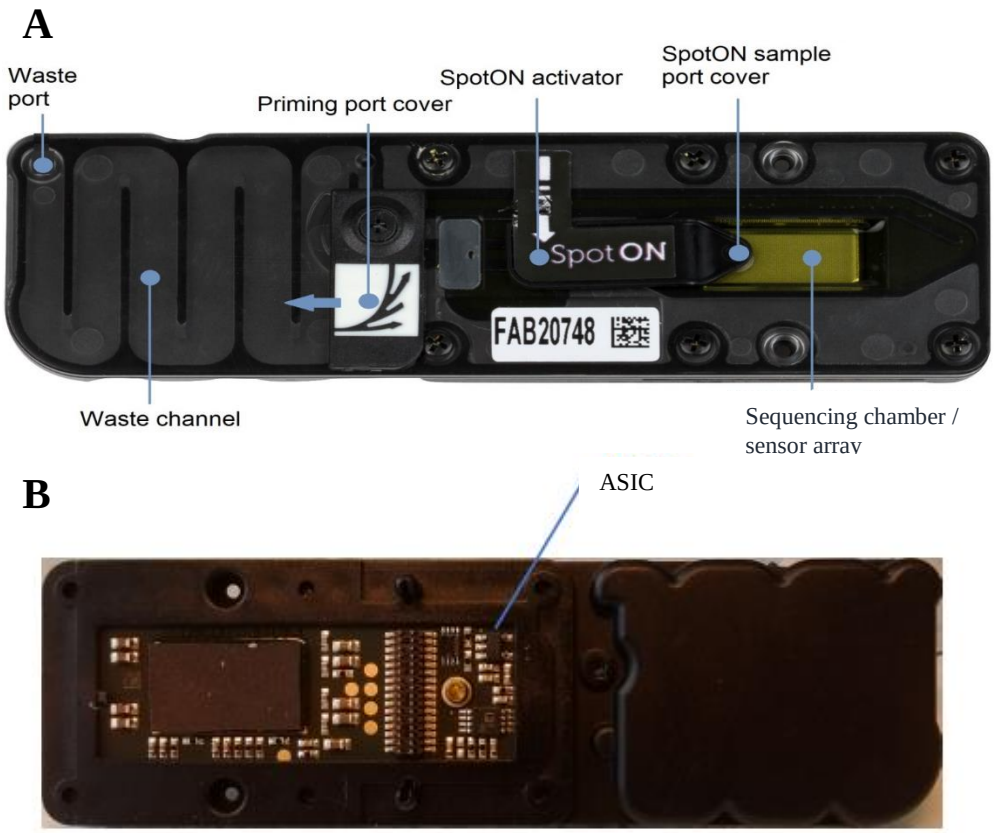

Image 5. The MinION flow cell, depicted as follows: A: the front view of the flow cell, B: the back of the flow cell.

The sequencing chamber, also referred to as sensor array thanks to the main functional element within it, is filled with electrochemically charged liquid. Inside the sequencing chamber lies a complicated and delicate structure of electrically resistant polypeptide membrane and 2,048 nanoscale pores penetrating said membrane. The membrane splits the chamber into top and bottom sections and upholds an electrochemical gradient between the two. This gradient over the membrane provides the electronic flow through the pores as it tries to achieve equilibrium, driven by controlled negative charge introduced by the electrical components of the flow cell during sequencing.

The Application-Specific Integrated Circuit (ASIC) on the bottom of the flow cell is a purely electronic component, fundamentally no different from microchips inside any modern computing device. As implied by the name, it has been specifically designed for the express purpose of controlling the electronic elements inside the MinION flow cell. The ASIC is both responsible for administering the voltage needed to control the electronic flow inside the sequencing chamber and processing the raw signal data

produced by the device.[12] Its operation is controlled through the MinKNOW software during runtime.

The ASIC chip has the capability of selectively controlling, activating or reversing the flow through single pores. The selective activation of pores is done automatically over the course of every sequencing run according to the state of the pores observed at the beginning of each sequencing run. At the beginning of the sequencing run the available pores are also categorized into four separate groups through a process called multiplexing. Initially, every sensor measuring the electronic flow has been assigned four separate pores of the flow cell to observe. Through the multiplexing process each sensor can effectively be focused on observing the most optimally performing pore throughout the sequencing experiment. The multiplexing also limits the number of simultaneously active pores to 512 from the total number of 2,048, helping to preserve the pore activity over the course of the run. By performing the multiplexing process again in the middle of the sequencing experiment, the more optimally performing pores can even be favored over less optimal ones, although for a regular run such step is not necessary. Overall effects of the multiplexing are increased yield and read quality.

The pore-specific temporal reversal of the electronic flow direction in turn is another important mechanism intended to increase the flow cell performance. It is used for unblocking and the possible reactivation of pores lost during the sequencing process. Pore loss can occur for various causes, such as stalling sample or unwanted impurities getting stuck inside the pore channels. Reversing the direction of the flow through the pores may help in removing these blocking factors and salvage the pore. The flow reversal is also performed automatically without input from the user by the MinKNOW software. For a period of time, the reversal was a global event performed to the entire flow cell at pre-determined time points during the run, but the newer MinKNOW versions have adapted pore- and situation-specific version of the process in an attempt to further minimize the amount of pore loss during the runtime.

2.2.3 The MinKNOW Sequencing Software

The MinKNOW software used to control the MinION sequencing runs is available for install to Windows, Linux and Mac environments and provides a Graphical User Interface (GUI) control option for the sequencing runs. The key feature of the software is that can be installed on a regular desktop or laptop computer instead of residing in its own dedicated control unit. This way the MinION sequencer remains untethered to a bulky and difficult-to-transport pre-made control unit configuration and allows the user to utilize the computing environment best suited for their sequencing environment to operate the device. The downsides of this arrangement are the additional setup required by the user to install and configure the software as well as possible additional costs associated with procuring a computer suitable for the sequencing experiment.

The MinKNOW software has gone through multiple iterations both during and after the experiments of this study. While the Graphical User Interface (GUI) and specific elements of the software have occasionally experienced quite drastic changes between different versions, the basic functions of the software have remained largely the same. All iterations of the software so far have contained two separate windows used for monitoring the sequencing run as it progresses, depicted in the Images 5 and 6. The Image 6 depicts the real-time pore status chart of the flow cell during the sequencing experiment. Each pore is represented by its own octagon with the color depicting the current status of the pore. The color meanings and pore layout in the graph have been re-defined over time as the result of updates to the software. The latest version of the MinKNOW software available at the time of writing (MinKNOW v. 1.13.1) has the pores organized reflecting their physical layout on the flow cell but the older versions used their own specialized layout not corresponding to the real physical location of the pores. The various status light and a detailed explanation of their meanings regarding the pore status for the original MinKNOW as well as the newest version have been listed in Table 1 for reference.

Image 6. The pore status view during active Nanopore MinION sequencing run in the original version of the MinKNOW software. The meaning of each of the possible pore status colors is explained in Table 3.

The other main view of the software is the yield graph illustrated in the Image 7. This view shows a visual bar graph representation of the reads produced during the current sequencing run segregated into suitable bins. Each bin depicts the total amount of bases sequenced in reads of similar lengths matching the label of the bin. Additionally, the global total number of sequenced bases is calculated and displayed at the top of the graph in real time, as well as a count of pores that have produced sequence data over the progression of the run. This screen can be used to assess the yield and progression of the run in real time and used to determine the completion of a pre-determined sequencing goal. If the target is achieved before the expiration time of the run the sequencing process can be stopped prematurely, possibly preserving some of the flow cell functionality for the later use.

Table 3. The meaning of pore status indicator colors in different MinKNOW versions.

Pore status	Color in the original MinKNOW	Color in the latest MinKNOW (v. 1.13.1)
The channel is saturated and does not produce sequence.	**Black**	**black**
The channel is passing very little current through.	**light blue**	**light blue**
The channel is passing current but has not been assigned for sequencing.	**dark blue**	**dark blue**
More than one pore have been detected by the sensor.	**Yellow**	**orange**
A single pore is detected and ready to produce sequence but no DNA strand is currently detected.	**light green**	**light green**
A strand is currently traversing (=being sequenced) by the pore.	**dark green**	**dark green**
The pore is blocked.	**Pink**	No assigned color
The pore is in its default state and has not been assigned any other status.	**olive brown**	No assigned color
An adapter is detected in the pore.	No assigned color	**peach**
The pore is unavailable and its status is unknown.	No assigned color	**light blue**
Active feedback.	No assigned color	**violet**

Image 7. The yield and read length histogram of active Nanopore MinION run as depicted in the original version of the MinKNOW software.

2.3 The Characteristics of MinION Sequencing

2.3.1 The Common Characteristics of MinION Nanopore Reads

The reads produced by the MinION are generally easily distinguished from the ones produced by the current market-leading technologies by their unique characteristics. As the sample material is not sequenced through chemical reactions but utilizing the unique physical shape of the nucleotides, many of the commonly accepted sequencing limitations do not apply to the MinION reads. The length of the sequenced target strands does not theoretically affect the nanopore sequencing process, effectively removing all technology-induced limits to the length of the reads. This theoretically speculated property of the nanopore sequencing has also been proven to hold true by the successful sequencing of reference-aligning ultra-long sequencing reads using MinION.[13, 14, 15] The lack of chemical reactions during sequencing run-time also means that nanopore sequencing is adaptable to various sample types, further discussed in chapter 2.3.2.

Even though the nanopore sequencing method itself poses very few limitations to the sequencing capabilities of the MinION, some do exist. Currently, the most prominent downside is the lower sequencing quality compared to its competitors. This can mostly be attributed to the challenges of signal measurement and processing. The applied measurement technology of today is not accurate enough to flawlessly capture all the minute changes of the electric current passing through the pores. The same applies to the basecalling algorithms that convert the raw signal data into sequence information. Traditionally, the homologous and repetitive genome regions with lots of proximal similarities are particularly challenging in this regard due to the very small signal changes caused as they pass through the pore. While improvement in this regard is achievable through improved structural and algorithmic construction, a certain level of uncertainty will always be inevitable. For example, the MinION and other nanopore sequencers will not be able to compete with the reaction-based sequencing methods in terms of pure scalability or sequencing output. This is caused by the more complex membrane-pore-sensor structure needed for the nanopore sequencers.

2.3.2 The MinION Sequencer Read Types

As previously mentioned, the nanopore sequencing technology can be relatively easily adapted to various sample types. The MinION device, for example, is currently being marketed for DNA, cDNA and RNA sequencing experiments. It has also been proven to work in distinguishing methylated nucleotides with relatively minor additional adjustments with all methods using the same flow cells and similar sequencing protocols.[16, 17] However, in this study all of the experiments were performed using only various DNA samples as there would have been no added benefit from utilizing multiple sample types for the building process of a DNA sequencing pipeline.

In addition, there are a few alternative options in the read types that are producible with the MinION. Depending on the library preparation protocol, the DNA can be sequenced either as a single-strand DNA or together as combined template and complementary double-strand complex connected by a hairpin adapter from one end. These two read types are known as 1D and 2D reads respectively. The Image 8 demonstrates the 2D structure of the sample during 2D sequencing and visualizes its traveling process through the pores.

The key difference between the two read types is whether the signal production and basecalling steps have the additional data from the complementary strand available to them. As the known sequence of the hairpin adapter connecting the two strands together can be recognized from read signal data, it can be used to combine the data from the complementary strands together for a consensus sequence. This gives every hairpin-connected read an intrinsic internal control against which the sequencing output can be compared to. On the other hand, the process of 2D sequencing is notably slower due to the slower passing speed and longer dwell time of the more rigid dual-strand complexes in comparison to the single-strand samples. This results in decreased yield per flow cell. The penetration process of the 2D strand configuration also includes some less-than-obvious challenges during the signal production. The movement speed of the 2D structured DNA fluctuates more than a simple single strand, the observed positive effect

of the consensus-based sequence generation is less pronounced than intuitively expected.

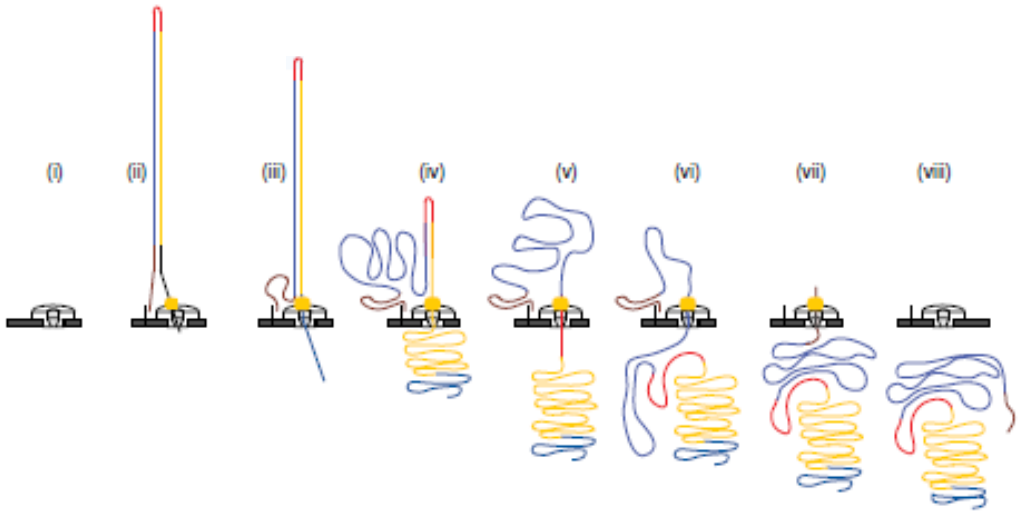

Image 8. A visual representation of the 2D sequencing process. (ONT)

Through the further development of the MinION, an additional read type called $1D^2$ was introduced. Having characteristics of both previous two read types, $1D^2$ reads can in a sense be considered to be a combination of the two. In $1D^2$ sequencing the template and complement reads are not structurally connected by a hairpin but instead share a double adapter ligated to both ends of the dsDNA. During sequencing, it is much more likely that these two complimentary strands go through the same pore immediately after one another instead of any other strand entering between them. The $1D^2$ sequencing protocol of the MinKNOW software looks for the characteristic markers indicating that the two complementary strands were indeed sequenced immediately after each other during signal production. If such markers are observed, the two signal measurements can then be processed in a similar way to the 2D sequences to increase the accuracy through consensus-calling methods.

Originally, the 2D sequencing was intended to become the standard sequencing protocol of the MinION. However, it was later withdrawn in favor of the 1D and $1D^2$ read types. This was done because of the improvements in the sequencing quality of the 1D and $1D^2$ read types and because of inferior yield and read counts of the 2D protocols. The

2D structure also sparked some legal dispute within the sequencing field as a competitor on the field, Pacific BioSciences, filed a copyright infringement claim against ONT. Although eventually resolved in favor of ONT, the removal of the 2D protocol from MinION eliminates any further legal trouble on the matter.[18, 19] At present the 1D and $1D^2$ sequencing methods are available while the production and the all products regarding 2D method have been discontinued.

2.3.3 The General MinION Library Preparation Principle

Regardless of the sample or read type, the basic steps of the MinION library preparation pipeline has remained unchanged. Thanks to this, the nanopore sequencing is readily adaptable to a heterogeneous selection of sample types with minor modifications to the library preparation and basecalling processes. In this sense, the actual characteristics of the sample are essentially irrelevant for the sequencing process itself as long as the sample can be effectively introduced to and passed through the pores. Since all the library preparation protocols only need to focus on successfully preparing the sample with this target in mind, regardless of the sample or read type itself, the general steps between them are bound to closely resemble each other.

The library preparation for the MinION run effectively consists of four stages: the sample preparation, processing of the strand ends, ligation of the sequencing adapters and preparing the final sample for loading into the flow cell. The purity of the library is ensured through multiple washing steps typically following each of these four main library preparation steps. However, the specific number of these wash steps vary between different versions of the library preparation protocol. Even though the general outline of the library preparation will be the same, certain sample types may demand slightly different handling or introduce some additional steps, such as the reverse transcription required for cDNA sequencing.

3 MATERIALS AND METHODS

3.1 The Sequencing Hardware

The focal point of this study was the MinION sequencing device from Oxford Nanopore Technologies (ONT). It is the first sequencing device developed, sold and produced by ONT MinION as well as the first commercially available option for nanopore sequencing. The MinION1 was first available in limited quantities through an application-based MinION Access Programme (MAP) before being released to open market a few years later.

The sequencing runs of this study were performed over the years 2014-2017 starting with the first version of the MinION sequencer received prior to the beginning of the study on 2014. The MinION device was replaced twice during this time period by a newer model upon their release. In addition to the device upgrades, the sequencing chemistry and flow cells both went over multiple upgrades during the progression of the study as well. The details of sequencing kits are elaborated on in section *4.2 DNA Sequencing Kit* while the detailed sequencing setup of each sequencing experiment of the study is documented in Table 4. Finally, the observed effect of the device upgrades over the course of the study to the sequencing results are discussed in the *Results* and *Discussion* sections.

For the duration of the sequencing, the MinION is connected to an external PC containing a pre-installed MinKNOW software using a provided USB 3.0 cable. During the first sequencing runs of this study, a Windows-based PC with the following internal specifications was used: Intel^{R} Core^{TM} i7-4770 3.40 GHz processor with 8GB RAM. After the Linux-compatible version of the sequencing software was released, it was replaced with a new Linux computer with the following specifications: Intel^{R} Core^{TM} i7-4790, 8 x 3.6 GHz processor and 8 GB RAM. The amount of RAM was later expanded to 12 GB to better handle the data analysis steps. The operating system on the

Linux computer was a University of Helsinki Common Ubuntu based Linux distribution Cubbli.

3.2 DNA Sample Material

3.2.1 Viral and Bacterial DNA

A Lambda DNA sample provided by ONT (DNA-CS, ONT, United Kingdom) was used in the study as a control DNA. It is a 3,6 kb standard amplicon that maps to the 3' end of the Lambda genome.[20] The genomic sequence of the DNA-CS can be found in the *Additional Data* section 2. The very first run using the MinION1 device was done with the control Lambda DNA sample provided by ONT to assess the correct operation of the device and test the sequencing protocol. This process is referred to as a burn-in run and was repeated whenever it was estimated to be necessary due to either dissatisfactory sequencing results or notable changes to the sequencing setup. The Lambda DNA was also used outside the burn-in runs as an optional internal control in the sequencing runs alongside the real sample as a way to better estimate the quality of the prepared library during some of the real sequencing runs. The genomic sequence of the DNA-CS can be found in the *Additional Data* section 1.

The control Lambda DNA was also used as a sample material in the basic performance calibration runs of the study as it was easily available and the small size of the genome simplifies the data analysis steps of the pipeline. Therefore, the Lambda was a well-suited sample material for the general sequencing pipeline optimization. As our pipeline construction process gradually moved towards more specific research questions the amount of helpful pre-existing documentation and available consultation consequently lessened as well. This necessitated the optimization and troubleshooting of the sequencing process through repeated experimentation. For such situations, the bacterial DNA was again used as sample material due to the aforementioned reasons.

Escherichia coli GST, a commercial *E.coli* strand with a *GST* plasmid (Thermo Fisher Scientific, US) was the bacterial DNA used in the study. For the express purpose of testing the MinION as a targeted sequencing method, some of the sequencing runs were performed using pGEX-4T1 plasmids with *NEB* exon inserts as the sample. Using this approach, it was possible to test the effectiveness of the targeting methods specifically towards our region of interest (ROI) while keeping the size of the total sequence pool much smaller compared to a whole human genome. The sequence of the plasmid and the inserted *NEB* exons are listed in the *Additional Data* section 1.

3.2.2 Mammalian DNA

The three human DNA samples were the main research interest of this study. The samples were patient samples used in accordance to the permissions of pre-collected concession forms. The DNA sample of Patient 1 had been previously received by our research group as an extracted DNA sample with no additional details regarding the used extraction method. The DNA sample of Patient 2 was extracted from cultured myoblasts using Gentra gDNA purification kit (QIAGEN, Germany). The Patient 1 sample was used as material in the Seq. run 1 and the Patient 2 sample in the Seq. run 2.

An additional human DNA control sample was obtained by extracting it from the blood sample of the writer following the extraction protocol by Lahiri and Nurnberger Jr. (1991).[21] This protocol was chosen because of its capability of producing very pure and non-fragmented DNA samples compared to the typical output from commercial purification kits. These sample qualities proved to be extremely important for the MinION method, necessitating the use of more traditional extraction methods. The chosen protocol was preferred over other options with similar sample qualities and output, such as traditional phenol-chloroform extraction, because it requires no toxic or hazardous materials. Furthermore, the protocol has been used by our research group successfully in the past to produce high-quality extractions of DNA for other experiments.

The purity and concentration of each DNA sample was measured with NanoDrop ND-1000 Spectrophotometer both directly after the DNA extraction and again immediately before starting the sequencing library preparation. The general length distribution of the samples was also confirmed using agarose gel electrophoresis whenever it was deemed necessary for the success of the experiment. The samples used for each of the sequencing runs are specified in the *Results* section and documented in Table 4.

3.3 Sequencing Library Preparation

All the sequencing libraries used in the study were prepared using the latest version of the recommended DNA sequencing kit or its immediate predecessor if such a kit was already available in our laboratory and still officially supported by ONT. Since the MinION device was under active development throughout this study, the sequencing kit recommendations changed often and the details of the library preparation protocol were in constant turmoil over the entire study period. As a result, the number of different sequencing kit versions used for the experiments of this study was uncharacteristically high for a single sequencing experimentation set.

Regardless of the multitude of the different sequencing kit versions used in the study the basic steps of the library preparation have remained comparably stable as stated earlier in the *Literature Review* section. The evolution of the library preparation necessitated a recurring replacement of old reagents or supplementary kits and adoption of new ones to cope with the newer sequencing kit versions. Such changes were entirely independent but often not chronologically separate from the changes to the experimental sample material.

As listing all the minor variations to the library preparation protocol over the experiment period would be out of the realm of possibility within this thesis, a detailed walkthrough of what could be considered the skeleton structure of the library preparation protocol is detailed below alongside a concept Image 9 visualizing the steps of the protocol. The more prominent changes within the protocol are discussed in the

relevant sections as appropriate. Additionally, all the versions of library preparation kits used alongside their version codes used during the study are documented in Table 2 and Table 4. Table 2 lists the overall changes to the sequencing setup and Table 4 documents every sequencing experiment of the study.

Table 2. The version codes of the sequencing kits, flow cells and library preparation kits used in sequencing experiments.

Run name	Product ID	Flow cell	Sequencing kit	Date of sample prep	Sample type
Burn-in 1	FLO-MAP001	MN-20-68571	SQK-MAP002	Pre-seq 03.07.2014, Final sample 07.07.2014	Lambda + Lambda CS
Burn-in 2	FLO-MAP001	MN-20-46821	SQK-MAP002	Pre-seq 09.07.2014, Final sample 10.07.2014	Lambda + Lambda CS
Burn-in 3	FLO-MAP002	MN-20-68543	SQK-MAP002	Pre-seq 17.12.2014, Final sample 18.12.2014	Lambda + Lambda CS
Burn-in 4	FLO-MAP002	MN-20-46627	SQK-MAP002	Pre-seq 29.01.2015, Final sample 30.01.2015	Lambda + Lambda CS
Burn-in 5	FLO-MAP002	MN-20-46627	SQK-MAP002	Pre-seq 29.01.2015, Final sample 02.02.2015	Lambda + Lambda CS
Burn-in 6	FLO-MAP003	FAA36042	SQK-MAP005	Pre-seq 23.04.2015, Final sample 24.04.2015	Lambda + Lambda CS
Burn-in 7	FLO-MAP003	FAA36042	SQK-MAP005	Pre-seq 23.04.2015, Final sample 24.04.2015	Lambda + Lambda CS
Burn-in 8	FLO-MAP003	FAA33770 (1st)	SQK-MAP005	Pre-seq 22.06.2015, Final sample 23.06.2015	Lambda + Lambda CS
Burn-in 9	FLO-MAP003	FAA33770 (2nd)	SQK-MAP005	Pre-seq 24.06.2015, Final sample 25.06.2015	Lambda + Lambda CS
Seq. run 1a	FLO-MAP003	N/A	SQK-MAP005	Pre-seq 07.09.2015, Final sample 07.09.2015	Human gDNA dilution + Lambda CS
Seq. run 1b	FLO-MAP003	N/A	SQK-MAP005	21.09.2015 (07.09.2015 run re-analyzed)	Human gDNA dilution + Lambda CS
Seq. run 2	FLO-MAP103	FAA64857	SQK-MAP006	Pre-seg 09.12.2015, Final sample 09.12.2015	Human gDNA dilution + Lambda CS
Seq. run 3	FLO-MAP103	FAA64626	SQK-MAP006	Pre-seg 30.03.2016, Final sample 30.03.2016	Nemalin myopathy gene probe sample + Lambda CS
Targ. run 1	FLO-MAP103	FAA64946	SQK-MAP006	Pre-seg 11.05.2016, Final sample 11.05.2016	Nemalin myopathy gene probe sample + Lambda CS
Targ. run 2	FLO-MIN104	FAD15215	SQK-MIN007	Pre-seq 20.07.2016, Final sample 20.07.2016	PCR NEB ex75, NEB ex77 & Acta1 ex5
Targ. run 3	FLO-MIN104	FAD15084	SQK-NSK007	Pre-seg 10.08.2016, Final sample 11.08.2016	PGEX4T-1 plasmids with NEB 54-1, 78-1, 122-1 and 151-1, ligated with XhoI
Targ. run 4	FLO-MIN104	FAD15084	SQK-NSK007	Pre-seg 10.08.2016, Final sample 11.08.2016	PGEX4T-1 plasmids with NEB 54-1, 78-1, 122-1 and 151-1, ligated with XhoI
Targ. run 5	FLO-MIN004	FAD15090	SQK-NSK007	Pre-seg 11.09.2016, Final sample 12.09.2016	PGEX4T-1 plasmids with NEB 54-1, 78-1, 122-1 and 151-1, ligated with XhoI
Targ. run 6	FLO-MIN004	FAD15090	SQK-NSK007	Pre-seg 25.09.2016, Final sample 26.09.2016	PGEX4T-1 plasmids with NEB 54-1, 78-1, 122-1 and 151-1, ligated with XhoI
Seq. run 4	FLO-MIN104	FAD24102	SQK-LSK108	Complete sample prep 7.12.2016	E.coli GST genome DNA, 1D sequencing kit & bead loading kit
Seq. run 5	FLO-MIN104	FAD24113	SQK-NSK007	Complete sample prep 14.12.2016	E.coli GST whole genome DNA, 2D sequencing kit

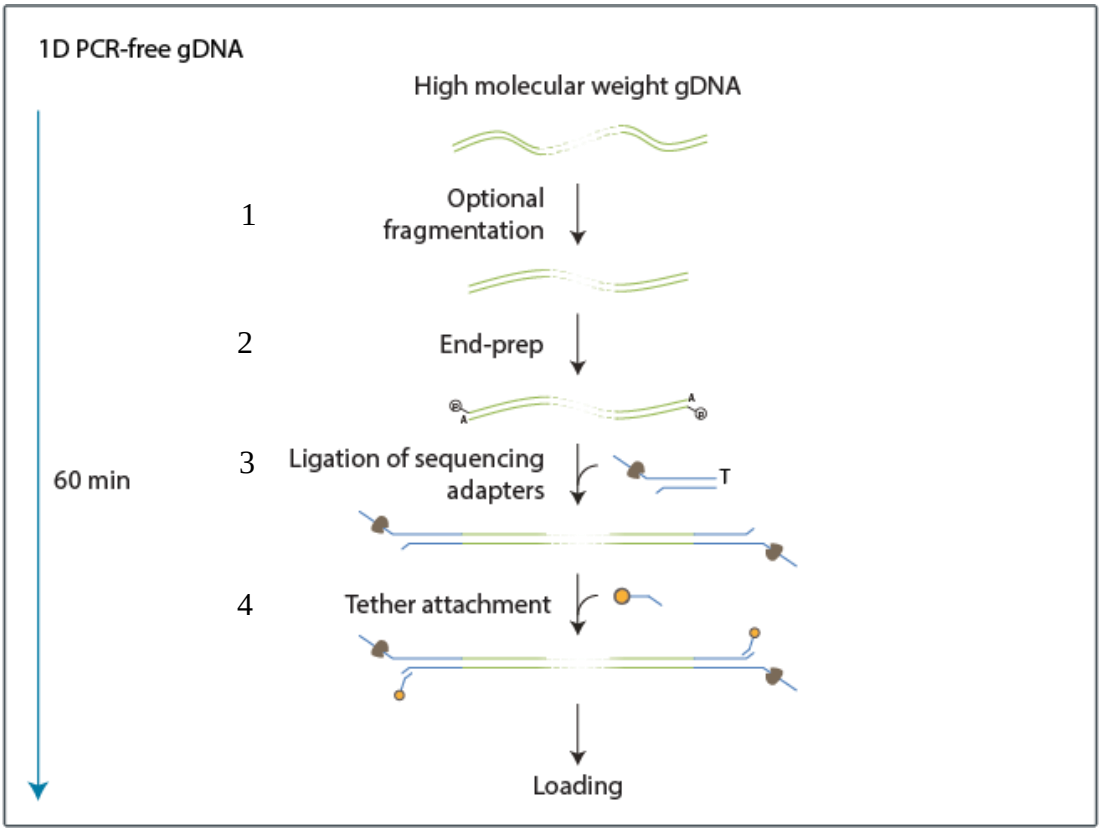

Image 9. The basic steps of the MinION 1D library preparation protocol. The following key steps are illustrated: 1: Optional fragmentation 2: End repair and dA-tailing of the sample fragments (End-prep) 3: Sequencing adapter ligation 4: Tether attachment. The washing steps following the End-prep, Adapter ligation and Tether attachment steps are not illustrated. The image has been modified from the ONT Library Preparation Protocol documentation.

The 1D ligation sequencing kit library preparation protocol is used as the basis of the provided protocol skeleton since it is both the current default protocol for MinION DNA sequencing and the simplest of all available sequencing protocols. The key steps of the protocol are illustrated in Image 9. Omitted from the image are the washing steps included in the library preparation.

The library preparation is initiated with an optional fragmentation step meant to ensure that the sequenced reads are of relatively uniform length. The early versions of the sequencing protocol were also innately designed to work optimally with DNA fragments of around 8 kb in length. The preferred fragmentation method for MinION is the g-Tube (Covaris, USA) that provides a quick and reagent-free fragmentation method for extracted DNA and does not require additional washing steps. To maximize the probability of success this fragmentation step was used in the early sequencing runs of this study. After the obtained yield from the sequencing experiments started showing

improvement this step was omitted from the pipeline. This was done since the long-term goal for the MinION was long-read sequencing, meaning the intentional fragmentation of the sample would be counterproductive.

The washing steps in several parts of the library preparation protocol refer to a simple bead wash of the sample with AMPure XP beads (Beckman Coulter, USA). The washing procedure remained the same for every version of the library preparation protocol during the study. First 0.7-1.5 times the sample volume of vortex-mixed beads were added to the sample and left to incubate in a hula mixer for 5 minutes in room temperature. Afterwards the sample was spun down and the tube was placed on a magnetic rack until clear pellet formation. Supernatant was removed from the tube by pipetting while still on the magnetic rack. The pellet was rinsed twice by adding and consequently removing 200 µl of fresh 70% ethanol directly to the tube. After a second ethanol wash the tube was briefly spun on a tabletop spinner, replaced on the magnetic rack and the residual ethanol was removed with a pipette. Then the pellet was suspended in nuclease-free water and incubated for few minutes in room temperature. Finally, the beads were pelleted again on the magnetic rack and the purified sample was recovered alongside the supernatant.

All the washes performed during the library preparations were done using this protocol except for the final wash of the library preparation. The early sequencing protocols used MyOne C1 streptavidin beads (Thermo Fisher Scientific, US) in their final wash. The newer protocols again omitted this extra requirement and the final wash was performed using the same AMPure XP beads as in the other washing steps. The only difference was that the final elution was not done using water but elution buffer (ONT, UK) from the corresponding library preparation kit.

The end preparation step of the sample has remained functionally unchanged over the period of the study. The end preparation entails two separate reactions performed on the sample DNA strands: an end repair step to first convert possibly any fragmented DNA strands into blunt-ended DNA and a dA-tailing reaction to attach poly-A tail to the ends

of these blunt-ended DNA strands. The first protocol utilized NEBNext End kit (New England Biolabs, USA) repair and NEBNext dA-tailing module (New England Biolabs, USA) to achieve these steps. This was eventually simplified through the implementation of a new replacement reaction kit combining these two reactions, NEBNext End repair / dA-tailing Module, (New England Biolabs, USA) in the later versions of the protocol. The only meaningful difference is the removal of the additional washing step that was included between the end repair and dA-tailing steps of the library preparation. However, the end products of these two kit configurations are otherwise identical. Therefore, most of the experiments in this study were performed using the two separate NEB kits even after they were officially replaced by the combination kit in the library preparation protocol as we had a pre-existing stock of them available.

Following the end preparation, the ends of the sample strands were ligated with the MinION-specific ligation adapters with the Blunt/TA Ligase Master Mix (New England Biolabs, USA) followed by a washing step. It was at this point the 2D sequencing protocol included the addition of hairpin adapters into the ligation reaction to attach the hairpin adapter to the ends of the dsDNA structure.

After the adapter ligation a tether component was added to the ends of the strands as depicted in Image 9. This tether is composed of cholesterol and it enhances the sequencing process by binding with the membranes of the channels inside the flow cell sequencing chamber. This binding naturally brings the DNA strands into close proximity of the pores. This enhances the entering of the sample strands into the pores during the sequencing process and leads into increased yield.

After the tether ligation the library was washed for the last time. This wash step differed from the previous ones in the original library protocols by using streptavidin beads and specific buffer and elution buffers instead of the AMPure XP beads. The current version of the protocol has replaced the streptavidin beads with the AMPure XP beads but still uses a proprietary elution buffer. The eluate is called pre-sequencing library mix and can be stored overnight in +4 °C. Typically this storage option was utilized when

reloading the MinION during sequencing run but was generally otherwise avoided during the experiments of this study.

Before the pre-sequencing mix was loaded into the flow cell it still had to be mixed with a separate running mix concoction to obtain the complete sequencing library. The components of this running mix are running buffer, (ONT, UK) fuel mix (ONT, UK) and nuclease-free water. For the experiments using SpotON-enabled flow cells another preparative step before the sample loading was introduced in the form of ONT Library Loading Kit. This kit contains loading beads which are mixed together with the sample immediately prior to the sample loading. First, the additional thickness and density of the complete library resulting from the loading beads makes the loading step of the sequencing library through the SpotON port much easier. Secondly, the beads have a similar function to the tethering adapter by pulling the sample material downwards towards the membrane and pores during the sequencing.

The sample loading process has been slightly altered in transition from the older flow cells to the newer SpotON flow cells. With the flow cell models without SpotON port the priming and sample loading were both performed through the same priming port shown in the image 5. Prior to the sample loading the flow cells were primed by pipetting priming mix through the priming port. The priming mix was obtaining by combining the three components added to the pre-sequencing mix without the sample. The priming process consisted of two loadings of 500 µl of priming mix, followed by 10 minutes of incubation after each one.

After the introduction of the SpotON port on the flow cells, the priming process became a bit more sophisticated. First 800 µl of priming mix was slowly pipetted into the flow cell through the flow cell priming port while keeping the SpotON port closed. Both ports were then closed and the flow cell was allowed to stand for 5 minutes. Next both the priming and SpotON ports were opened and 200 µl of priming mix was pipetted through the priming port in a continuous motion. At this point a slight liquid overflow was observed from the SpotON port, indicating that the flow cell was ready for the

sample loading. The sequencing library sample pre-mixed with the loading beads was pipetted into the SpotON port slowly drop-by-drop using regular laboratory pipette. Finally, both ports were covered and the sequencing run could be started.

The applicability of the flow cell washing protocol and their re-usage for multiple sequencing experiments was also considered in this study. The impact that a single sequencing run had to the state of the flow cell proved to be highly inconsistent, with some flow cells showing very little signs of wear while others were almost entirely expended after a single sequencing experiment. While such variations are partially attributable to factors such as the original quality of samples and difficulties during library preparation, some were undeniably caused by factors outside of the user's control. Such intrinsic factors include the poor quality of the individual flow cells and non-user related problems in sample loading, among others. To mitigate the loss of resources caused by such matters the flow cell washing protocol was tested as an attempt to either re-use a once run flow cell again or to preserve ones left virtually unused due to unsuccessful library preparation. The preservative effectiveness of the washing to the quality of the flow cell was however observed to be both limited and inconsistent. The time a washed flow cell could maintain a reasonable pore activity was observed to be only a few weeks. This was of limited use for our study, where the time gap between runs could become notably large depending on the availability of the samples, the success of the previous run and the updates made by ONT during the interim periods. The sequencing objective of this study was strongly geared towards attempting to maximize the sequence yield during a single run of a sample. The washing protocol was consequently determined not to be particularly beneficial for this study and the flow cells were generally considered single experiment consumables.

3.4 Target enrichment

After establishing the basic operations of the MinION device, the possible integration of target enrichment into the sequencing protocol was investigated. The enrichment method selected for these experiments was xGen® Lockdown® Probes by IDT (Integrated DNA Technologies, US). For this purpose a capture probe array comprising all ten nemaline myopathy genes published at the time: *ACTA1*, *CFL2*, *KBTBD13*, *KLHL40*, *KLHL41*, *LMOD3*, *NEB*, *TNNT1*, *TPM2* and *TPM3*. The gene *YBX3* was also included in the probe set. At the time of probe design this gene was one of the potential genes of interest in nemaline myopathy research and was later associated with the disease.[22] The design for the probes were done using the IDT Target Capture Probe Design & Ordering Tool (Integrated DNA Technologies, US).[23]

The Lockdown probes were used according to the IDT xGen® Lockdown® Probe Hybridization capture protocol version 1. The probe solution was thawed at room temperature, mixed and spun down using eppendorf 5424 tabletop centrifuge.

3.5 Sequencing Software

All sequencing runs were operated through the proprietary MinKNOW software provided by ONT for the express purpose of MinION sequencing. Like the sequencing kits, the MinKNOW software went through multiple updates over the progression of this study. The MinKNOWsoftware also often received a new graphical look alongside its algorithmic changes. However, the main functions of the sequencing software remained mostly the same. Even as the newer versions of the program changed the visible GUI and gave the users easier access to controlling the specific elements of the sequencing run, most of the operations have been accessible indirectly through additional scripting solutions since the early versions of the software. The three major visual overhauls of the MinKNOW software over the course of the study are depicted in Image 10. The changes in the graphical look of the MinKNOW software may also be observed in the variance between the visual documentations of different runs.

While the effects of GUI changes to the progression of the sequencing runs are negligible, the underlying changes to the basecalling process of the raw data over the past years cannot be left unaddressed. The first available basecallers employed event-based basecalling algorithm where the raw signal data was segmented into smaller sections called events. These events were then assigned a kmer sequence corresponding to the closest matching estimate from the internal models of the basecaller. In essence, each event was compared to the internal database of the basecaller to find out which kmer had the most similar event profile. The exact length of these kmers was a strongly discussed area of development during this study, but for the most part the early basecalling solutions used 5mers in their algorithms. The basecaller originally available for this process, Metrichor, was a server-based solution with strict online connection requirements. This basecalling procedure entailed uploading of the raw sequencing data to the ONT-moderated cloud using a windows desktop application where it was then analyzed. The original files were then augmented with the basecalled sequence data and downloaded back to the local environment. This process was read-specific and could be performed simultaneously with the sequencing, allowing the produced sequence data to be basecalled nearly in real-time. The overall statistics of the basecalling result were available for review on the web-based user interface of the Metrichor and could also be compiled into basecalling reports.

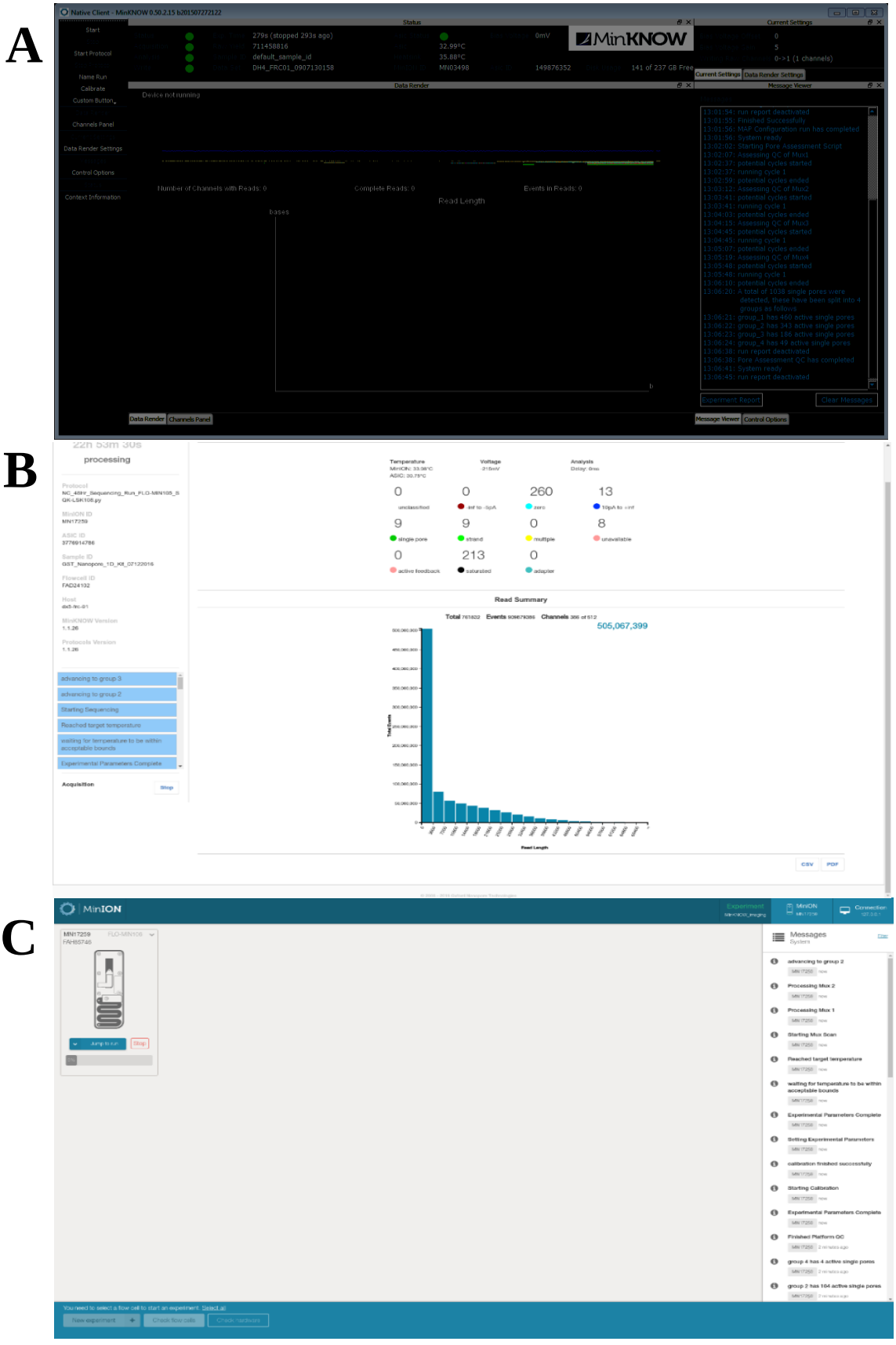

Image 10. The three distinct graphical looks of the MinKNOW software, from the oldest (A) to the newest (C).

3.6 Basecalling Software

The basecalling process has experienced three large-scale changes after the original Metrichor basecalling process. The first was an upgrade to the original Metrichor basecalling software algorithm that expanded the kmer length used by the basecalling algorithm in event matching. The original version of the Metrichor basecaller matched the signal event data to its internal kmer segments of five nucleotides of length. However, the 5mer length of the algorithm caused notable difficulties when handling longer and more homogenous sections. This approach also had rather high intrinsic error rate simply because of the uncertainty of the event calling and matching steps. These issues were partially remedied through the extension of one additional nucleotide into the matching algorithm, leading into the event matching to be performed on the sections of six nucleotides instead of five. This helped the matching algorithm to handle the repetitive regions better. It also improved the overall accuracy of the basecalling through a more detailed event matching made possible thanks to the exponentially increased number of matching patterns. However, the overall effects of this change were still relatively minor. The most powerful and advanced analysis tools generally started to move away from event-based analysis and focused on utilizing the raw signal data instead.

The second and much more significant update to the basecalling protocol was the conversion from the event-based basecalling into Recurrent Neural Network (RNN) based algorithm, changing the fundamental principle of the basecalling. The modern basecallers have moved more and more towards RNN data structure in assigning the basecalls based on the raw signal data directly without the intermittent event-calling step. This algorithm was adapted in order to circumvent the fundamental limitations and data loss by the event distribution on the fluctuating signal data. This contributed in achieving higher basecalling accuracy. The detailed principles and operation of these basecalling algorithms and data models is outside the scope of this thesis but their observable effects on the sequencing results are taken into account in result analysis.

The third notable paradigm shift in the basecalling process happened in parallel with, and partially because of the modifications to the algorithms. The original one-track basecalling pipeline branched out into multiple alternative directions for various cases. The previous basecalling standard Metrichor was replaced by a new local successor named Albacore. Albacore implemented a locally operated RNN-based basecalling algorithm and effectively removed the necessity of an active internet connection during the basecalling process. Additionally, the MinKNOW sequencing software was also expanded with its own basecalling functionality to preserve the option of live basecalling during sequencing, something that Albacore was not capable of. A third official basecalling option is the Guppy software (ONT, UK). Unlike the other mentioned basecallers, Guppy is specifically targeted to the people planning on developing or modifying the basecallers themselves and is also available through a specific developer license agreement. Recently, some third-party basecalling options have also become available, though none of them were implemented in this study. The Metrichor was used for the early stage basecalling as long as it was supported. After the release of the Albacore, it was adopted as the new standard. Some of the earlier sequencing runs were also re-basecalled using the newer algorithms.

3.7 The Bioinformatics Environment

The bioinformatic analysis of the sequencing data was performed partially locally in the same Linux desktop environment hosting Linux version of MinKNOW and partially in the high-performance computing environment Taito-Shell managed by CSC IT Center for Science, referred to henceforth as Taito-Shell and CSC.[24, 25] All the file trafficking between the local environment and Taito-Shell was done using Linux-native file transfer *scp* command to transfer the files securely over SSH connection. All the key programs used for the data analysis are freely available for scientific use from their respective download pages, which will be referenced at the introduction of each program. In addition to the available free programs the complete analysis pipeline was augmented with self-written support scripts used for file management, conversion and selective data extraction. These scripts are referred to in this thesis but their complete codes have been excluded due to their basic-level functionality and easy reproducibility.

3.8 Data analysis

The basic steps of the data analysis were established early and they remained stable over the entire study, although the specific program configuration was updated many times over the progression of the study. The basic flow of the analysis pipeline was:

1. Basecalling the sequence data and converting it into standard sequence file types
2. Combining, formatting and quality checking the produced sequence files
3. Aligning the sequence data against a reference genome with a suitable genomic alignment tool
4. Sorting, formatting and quality checking the alignment files
5. Visually analyzing the alignments using an alignment visualization software
6. Compiling statistic results of the sequencing run

In the early stages of the study the analysis pipeline was intentionally kept extremely simple and any additional filtering or quality assurance steps were excluded. This was done in order to compensate for the low yield and read quality of the early sequencing runs and to maximize the amount of data used for the analysis. When the performance of the sequencing runs improved and the basic structure of the analysis pipeline had been established, additional filtering steps were included. The aim for the final pipeline was to offer functionality and reliability comparable to the ones used in conventional sequencing studies.

The data analysis was always started by performing basecalling of the raw sequencing signal datasets. The raw data obtained from the sequencing was stored as fast5 files, a derivative of HDF5 file format. The sequence data was added into these files and extracted into more common standard *.fasta* and *.fastq* file formats during the

basecalling process. The Metrichor was the original basecaller used for the basecalling while the data extraction was achieved with poretools and PoRE R package utilities. [26, 27, 28] As these programs were only used for the extraction of the sequences from the fast5 files, their output is fully interchangeable. The commands used for the process are well documented in the manuals of the programs.[27, 28]

Immediately following basecalling, all reads from the sequencing run were pooled together into a single fastq file with a simple Linux environment command:

$$cat\ *.fastq\ >\ pooled\_reads.fastq$$

In certain instances, the resulting file was found to contain formatting errors as a result of the residual line breaks or similar additions that had been inserted into the single-read sequence files by the file conversion tool. In such cases the formatting of the pooled sequence file was fixed by removing empty lines with a *sed* command:

$$sed\ -i\ '/\^\backslash s*\$/d'\ file.fastq$$

In the case of particularly ambiguous formatting errors the problematic lines were removed from the sequence file with *sed* using the following command structure:

$$sed\ -e\ '5{,}10d;12d'\ sequence\_file\ >\ new\_file$$

The quality and basic properties of the reads were then checked using freely available tools FastQC and fastqp.[29, 30] FastQC was chosen as the default tool for quality comparisons while the fastqp was reserved as a supplementary option.

The fastq files of pooled reads were aligned against reference genomes using suitable alignment programs. The original reference used for the human DNA reads was GRCh37 which was later replaced by the newer GRCh38 release. The bacterial reference for the *E. coli GST* has been described in the

DNA Sample Material section 3.2.1, and the sequences of the plasmid/exon inserts are included in the *Additional data* section 2. The early sequencing runs of the study were aligned in the local Linux environment. However, as the yield of the sequencing experiments increased this became too time-consuming and the alignment steps were moved into the Taito-Shell as mentioned previously in the *Bioinformatics Environment* section 3.7.

The program used for the alignment production was changed multiple times over the study as various software developers continued developing alternative options better suited for the processing of the unique Nanopore reads. The original aligner used was LAST, recommended by ONT at the early stages of MAP due to its general focus towards long alignments.[31, 32] The LAST alignment pipeline was parallelized with the help of *parallel-fasta* command and the alignment was performed using *lastal* command. The completed alignment was temporarily stored as a *.txt* file before being converted into a sorted bam file with the help of the *maf-convert* utility. The command structure for the used LAST alignment pipeline is available in the *Additional Data* section 3.

The LAST aligner was replaced relatively early in the study by other alignment options that offered simpler and more reliable alignment pipelines, superior alignment quality and direct support for MinION read alignment. Therefore, it was eventually dropped from the list of alternative alignment programs altogether. The main aligner used for most of the study were Burrows-Wheeler Aligner (BWA), more specifically the *BWA-MEM* algorithm supported by additional alignment option Graphmap. [33, 34,35] The command pipelines used in this study for these aligners can be found from *Additional Data* section 3.

To assess the results of the sequencing experiments the *.sam* alignment files were converted into a more size-efficient *.bam* file format, sorted by genomic position of the alignments and indexed using samtools software utility.[36, 37] The commands used to achieve this are listed in the *Additional Data* section 3. For alignments performed in the

Taito-Shell the complete alignments and their index files were downloaded back into the local environment for the quality assurance and result analysis steps. The general statistics of the alignments were collected from the bam flagstat data field with the samtools utility by using a Linux terminal command:

$$samtools\ flasgstat\ Alignments.sorted.bam\ >\ Alignments\_flagstat.txt$$

More detailed quality data of the alignments was collected with the BamQC program.[38] The alignment coverage statistics were also separately collected using *bedtools* and visualized graphically by using basic drawing functions of R. The alignments were visually inspected with the help of two separate alignment visualization programs, Interactive Genome Viewer IGV and BasePlayer. [39, 40, 41, 42, 43] The BasePlayer visualizer had superior performance when viewing large genomic regions of alignments and was used to observe the overall depth and positioning of the alignments using its dynamic read-depth graph. After the regions of interest were identified from the alignments in BasePlayer, they were examined at the nucleotide-level using IGV.

The analysis pipeline was eventually strengthened through additional filtering and data selection scripts to obtain answers for more specific research questions regarding the alignments. Basic filtering operations included removing secondary alignments and unaligned reads or filtering our alignments below certain quality threshold, using *samtools view* command selectively. If the sequencing experiment had been performed including the internal Lambda DNA CS control, the control reads could be removed from total pool by a two-step process. First, all sequences were aligned against the Lambda DNA CS reference and the names of strongly mapping reads were gathered based on the results of this alignment. The removal of these control reads from the original pooled sequence file was then done using the in-house java script *mappedReadRemover.sh*. The script accepts a list of read names for removal and a multiread fastq file as its input. It then uses basic string matching and line removal commands to write a new fastq file containing only the reads not included in the given list of read names.

The more specific research questions included listing the alignment coverage by reference exons or separating the single longest aligned read from the total alignments file. This longest read was then used as the input for online Blast tool to confirm the reliability of the maximal length alignment of the dataset. [44] The exon-specific coverage was produced by first downloading exon features from UCSC Table Browser and then processing the data using *awk* utility into tab-separated Feature -bed file with the following features:

$$ChromnosomeFeature\_startFeature\_endExon\_numberFeature\_strandedness$$

This file was then used as the input alongside the alignment file for the *bedtools* command:

$$bedtools\ coverage\ -b\ Alignments.sorted.bam\ -a\ Features.bed\ >\ Coverage.txt$$

Extraction of the longest aligning read was done by combining *samtools view* and the native Linux *awk* file processing commands. The alignments were first displayed with the *samtools* view and piped as input to *awk*. The read names and alignment lengths were then filtered from the data and sorted with in-build *awk* functions and printed into a new file. This output file contained a list of all aligned reads in the order of their alignment length.

In addition to manual filtering steps, the MinKNOW software received its own intrinsic quality filtering at fairly early stage of the study. This internal quality filtering segregated the produced raw reads into two folders, named Pass and Fail, according to ernally defined parameters. For the 2D sequencing experiments, the main parameter used for this split was the length difference between the template and complement strands of the read. For the 1D and newer $1D^2$ technologies, the defining parameter was instead changed into a more general average sequencing quality as estimated by the MinKNOW software during sequencing. For the sake of analysis pipeline completion and robustness the practical steps needed to combine these two datasets were explored and tested in practice. Only the reads from the Pass folder were used in this study unless explicitly mentioned otherwise.

4 RESULTS

4.1 The Burn-in Experiments

The basic operations for the sequencing pipeline were established through standardized protocol burn-in runs before pursuing sequencing of actual research samples. As established in the *Materials and Methods* section, these burn-in runs used Lambda DNA-CS (ONT) as the sample material. This allowed the sequencing protocol to be performed with a minimum amount of variation, which is exceedingly important when establishing the baseline functionality. All the burn-in experiments and their key characteristics are listed in Table 4. These key characteristics include properties such as the sample type, the dates of library preparation and sequencing run as well as the specifics of the used library preparation protocol. Each run has also been assigned a unique run name for reference in this thesis.

The first burn-in experiments, Burn-in 1 and Burn-in 2, used the 2D sequencing kit and the early MAP release versions of the MinION flow cells. The sequencing and basecalling were done on a Windows computer using MinKNOW and Metrichor programs respectively, followed by the data analysis in the Linux environment. These sequencing experiments produced in total 2,742 and 188 reads respectively with none of them passing the Metrichor quality check. A small portion of these reads could be successfully aligned against the reference but the coverage and depth of these alignments were insufficient for any meaningful analysis.

Considering the poor initial results of the first burn-in runs we deemed the MinION unfit for processing any real samples at its current level of performance. Instead, a second set of burn-in experiments was performed in an attempt to improve the results. These are the runs Burn-in 3 and Burn-in 4 (Table 4). This second patch of burn-in runs utilized the newer FLO-MAP002 flow cells released after the first experiments. In order to address the lacking performance of the earlier runs the sequencing protocol itself was also adjusted with minor changes such as slightly increased incubation times and more

thorough mixing of reaction reagents and reaction mixes.The results from the second set of burn-ins were a notable step down compared to the first set, producing only a miniscule amounts of 8 reads for the first run and 206 for the second. Additionally, none of the reads passed the Metrichor quality filtering. Before any further experimentation, the possible causes for these low yields were researched from literature and the community resources provided by ONT. Our hypothesis and conclusions regarding these causes will be elaborated on in the *Discussion.*

During this time the MinION flow cells received another production upgrade and sequencing kit experienced multiple updates from version SQK-MAP002 to SQK-MAP005. The notable methodological updates and improved understanding of the protocol were deemed to be expansive enough to warrant a re-evaluation of the MinION performance through new burn-in experiments. The immediate action of improving the sterility of the work environment by switching to the use of filtered pipette tips and pre-emptive sterilization of work areas was also incorporated into the library preparation to try to improve the results of the previous burn-in attempts.

The Burn-in 6 and Burn-in 7 were the first burn-in runs using the upgraded flow cell and sequencing kit. The two runs were done using a single library and flow cell, so their output is handled as a single entity. The burn-in produced again a low yield of 187 reads. However, in sharp contrast to the previous attempts, over half of the sequenced bases passed the quality filtering of the Metrichor. Upon further analysis, the overall quality of these sequences was still observed to be low with the highest read accuracy reported by the Metrichor being 67%. However, a notable portion of the produced reads passing the Metrichor quality check was a promising improvement to the overall performance.

The next set of burn-ins, Burn-in 8 and Burn-in 9, were performed on a single flow cell on two separate days. Each burn-in experiment also had its own separate sequencing library but used the same flow cell. The flow cell was washed between the two runs using the early version of the MinION Flow Cell wash protocol available.

The Burn-in 8 produced 638 reads. The Burn-in 9 was performed after washing flow cell and produced 227 reads. In both cases a majority of the sequenced bases were within the reads that had passed the Metrichor filtering. The maximum quality of the reads from these runs also experienced a notable increase to respective values of 86% and 85%. All of the burn-in experiments using the FLO-MAP003 flow cells had also yielded sequences that passed the Metrichor quality filtering while exceeding the total length of 10 kb. This remained constant even with the burn-in library preparation protocol including a forced fragmentation step to around 8 kb medium fragment length.

It was clear that the updated burn-in protocols were performing much better and producing more consistent results than the previous versions. However, these results were still not nearly good enough for the MinION to be a practical tool for sequencing human samples in any real capacity. However, with the establishment of the basic sequencing protocol addressing these issues could be simply a matter of optimization, which the following experiments were focused on.

Since the focus of the pipeline development was switched from establishment to sample-specific optimization, the sample material for the following experiments was changed from the provided Lambda DNA-CS to bacterial and human DNA samples. The reasoning behind this change was the expectation that any successful performance enhancements achieved with additional Lambda DNA-CS burn-in experiments would have to be re-evaluated and modified to accommodate for the characteristics of real sample DNA. Additionally, this opened up a possibility of selectively spiking the future sequencing experiments with additional DNA-CS in order to assess whether the fluctuations in the observed sequencing throughput were sample- or protocol specific. This type of internal control was used when deemed necessary during different optimization runs.

4.2 The Sequencing Experiments

The first experiment following the concluded burn-in experiments, Seq. run 1a, was designed using a pre-existing sample of human genomic DNA. The goal of this experiment was to establish the performance baseline for the sequencing experiments outside the Lambda DNA samples. As described earlier, the Lambda DNA-CS spike-in was however included during the library preparation. The protocol implemented was the same that was used in Burn-in runs 8 and 9 with a minor change in the final steps of library loading and sequencing run initiation. While the pre-sequencing mix had previously been stored at fridge temperature overnight before the initiation of the sequencing, it was instead immediately processed to completion and the library was loaded to the sequencer without delay, after which the sequencing process was initiated.

This first sequencing run (Seq. Run 1a) was performed on 7th September 2015. The sequencing process was allowed to continue for nearly 40 hours. However, sometime during the last third of this time period the MinKNOW software encountered an error and crashed. The system was manually restored and restarted after an estimated downtime of 6 hours. Afterwards, the sequencing was resumed for an additional 8 hours. The overall yield from the run was 9,737 reads comprising 5 Mbases of sequence, out of which 689 reads passed the Metrichor filtering. The highest 2D quality score reported by Metrichor was 9.3 on the Phred scale, translating to an approximate base accuracy of 88%.

The same dataset from the first sequencing experiment was later run through the updated Metrichor basecaller again on the 21st of September, 2015 (Seq. run 1b), i.e. the existing data was a simply re-basecalled with a newer algorithm. The second analysis shifted the results of the run slightly with the new yield statistics being 9,951 reads (5 Mbases), 681 passing the quality filter and the highest reported accuracy value being 8.1 on the Phred score, translating to roughly 84% accuracy. On both versions of the dataset a majority of the reads passing the Metrichor filtering could be successfully aligned against the human reference genome using LAST aligner. The Metrichor-filtered reads

not aligning to the reference made up a fraction of around 10% of the total read count. In this group were also included the reads mapping against the DNA-CS reference.

If the sequencing results from the latest burn-in experiment (Burn-in 9) were to be extrapolated over the same run time as the human sequencing experiment, the yield expectancy from it would have been be around half of what was observed in real experiment. This confirms that the combination of our modifications to the protocol, improved familiarity with the system and different sample material had resulted into a notable increase in the sequencing efficiency. However, with the observed yield of 5 Mbases, only around 0.15% of the whole human genome could be expected to be covered once by a single sequencing experiment. For the MinION to be a viable option for sequencing of human samples the yield of single sequencing experiments would have to improve significantly. Another option would be to implement a targeting protocol into the sequencing pipeline and focus each sequencing experiment towards only a few specific regions of interest.

Before looking into the possibility of targeted MinION sequencing, we compared our observed sequencing output to the results obtained by other users and developers of the device. We were aware that the yields of both our burn-in and the sequencing experiments had been drastically below the theoretical capacity of a single MinION flow cell and sequencing run. Through further research we also confirmed that the yields reported by ONT and some of the other MAP participants were also much greater than what we had been able to achieve. While this did prove that further improvements to the sequencing protocol were indeed possible, it also confirmed that our understanding of the sequencing process was still insufficient. In reaction to this observation the optimization initiatives towards increased the amount of data obtained from a single flow cell and target enrichment were launched.

The first goal was to further increase the yield. We started by improving the quality of the starting material and switched to a fresh DNA sample extracted from cell pellets as described in the *Materials and Methods* section. This new sample was first tested on a new SQK-MAP006 version of the 2D sequencing library preparation kit. This experiment (Seq. run 2) produced 48,769 reads over the course of circa 48 hours, corresponding to 45 Mbases of sequence data. Furthermore, all of the reads passed the preliminary quality control filtering by the Metrichor basecaller. Finally, this run also produced the longest single read of 127 kb.

Despite the overall improvements we had been able to achieve, the total sequencing capacity of the MinION was still far below what would be required for an entire human genome sequencing experiment. We estimated that while the sequencing capacity of the MinION was only likely to increase further and further going forward, it was not likely to reach the required levels in the near future. Additionally, with reliable targeting method our current sequencing output could be robust enough to achieve our primary goal with the device. With these considerations in mind, our next experiments were directed towards building a complementary target enrichment protocol to accompany the MinION sequencing. Our next two experiments (Targ. run 1, Targ. run 2) were focused on testing such a target enrichment method. The sample for these experiments was again the same fresh DNA extraction as in the previous run. However, this time the sample had been pre-processed before the MinION library preparation with the xGEN lockdown probe protocol in an attempt to extract the ROIs from the total genome. Both runs produced a fairly low number of reads, 4,049 and 4,321 respectively. While some level of decrease in the read counts could be expected, the total read counts falling below those obtained from the Seq. run 1 was still disappointing. A drop in the maximum read length of the run was also observed, with the longest filtered read being around 21 kb of length. On the other hand all the reads passed the Metrichor filtering, which was still a notably better result than what the early experiments had produced. The majority of the reads also seemed to align near the expected target region of our reference genome, although due to their low count we hesitated to draw any conclusive results about the reliability of the targeting method. The results we obtained from the next experiment performed on 11.5.2016 (Targ. run 2) using the same experiment setup

were remarkably similar with every performance measurement statistic strongly mirroring the previous experiment.

In order to assess these observed poor read counts of the first target enrichment experiment we decided to adopt a new approach for the targeting experiment. Our hypothesis was that our protocol was functional on the basic level, but it was not capable of producing enough high-quality targeted DNA for our sequencing library preparation. We speculated that this could be related to the size of the target region compared to the entire genomic sample, as well as the length of our target fragments. The relatively small size of the ROI could make it difficult for the probes to locate and successfully bind with their complementary regions. Our goal of long-length fragments could further push this problem as the long DNA strands present in our sample could possibly become tangled or change its shape in such a way that the ROI could become inaccessible to probes. Finally, even if the probes could successfully attach to their complementary region, the DNA fragments could either tear or become too strongly bound due to their length, causing them to either be washed away or not be eluted into the sequencing sample. These factors could also help to explain the lower maximum read length we observed in our targeted experiments, as fragments of short-to-moderate length could possibly evade many of these factors.

Our next experiments were designed to test this hypothesis. We repeated the previous sequencing experiment this time using nebulin minigene insert plasmids as the sample material. Compared to our previous targeting experiment with the minigenes, this time the targeted regions would comprise a much larger total proportion of the sample fragments. The overall length of the fragments present in the sample material should also be much smaller compared to a whole human gDNA sample. If our hypothesis regarding the targeting protocol was correct, we should be able to observe improved results from this experiment composition compared to the previous tests.

The first experiment using the plasmid DNA was performed on July 20th and repeated on August 11th 2016. These are the listed as Targ. run 3 and Targ. run 4, consecutively.

The Targ. run 3 produced just over 53,000 reads, out of which almost 67% (35,630) passed the Metrichor filters. This corresponds to the sequencing output of 19 Mb. The Targ. run 4 also utilized a new experimental Metrichor basecalling algorithm including an integrated sequence matching against the human genome. It yielded 217 Mb of sequencing data in 64,630 reads with the passing count of 34,493 reads. Running the raw signal data of the Targ. seq 4 through the regular basecalling algorithm further increased the yield to 260 Mb and the read counts to 86,188 total, 44,997 after the filtering.

As we suspected, the targeting protocol worked much better on a plasmid-based experiment setup compared to the entire genome sample. We repeated the experiment setup with two more plasmid samples both to confirm our findings and to test if we would be able to push our results further with minor tweaks to the targeting process but were unable to achieve any significant improvements in the sequencing result. We eventually concluded our targeting experiments with the xGEN probe protocol on the Targ. run 6, which is the final experiment regarding the targeted sequencing discussed in this thesis.

4.3 The SQK-LSK108 protocol experiment

After the conclusion of our targeting experiments the MinION received two rather notable updates. A new SQK-LSK108 sequencing kit was released for both the 2D and 1D protocols, promising notable improvements both in the sequencing yield and stability. Additionally, the basecalling procedure experienced a large overhaul, the announced impending retirement of the cloud-based Metrichor service. In preparation for this, a local basecaller named Albacore was made available for the general users of the MinION. This change was made both in preparation to the expected increase in the volumes of basecalled data as both the average yield and the size of the MinION userbase continued to increase. The Albacore basecaller was also supposedly closer to the internal basecalling algorithms of ONT in performance compared to the Metrichor, which should translate to increased basecalling speed and accuracy for the end users.

We planned and executed a standard sequencing experiments for both the 1D and 2D library preparation kits to assess these supposed improvements. These experiments were performed using E.coli GST DNA as the sample, and the local Albacore basecaller as a new standard of the sequencing pipeline. The 1D experiment Seq. run 4 was performed on 7.12.2016 and the 2D experiment Seq. run 5 on 14.12.2016. However, after our 2D experiment, ONT completely removed any support for 2D sequencing technology from all their products with no apparent plans for its future reintroduction. Instead, a new $1D^2$ method offering similar benefits was introduced as a new option. The principles of this new $1D^2$ technology have been explained earlier, but a more detailed performance evaluation is out of the scope of this thesis. Consequently, since 2D technology has been deprecated, the more specific results of the Seq. run 5 will not be discussed and the following result analysis will be strictly focused on the Seq. run 4 results instead.

Seq. run 4 produced over 600,000 reads with over 90% mapping rate on default settings. Post-alignment average coverage over the whole E.coli genome was around 180. The coverage distribution over the reference genome has been depicted in Image 11. The mean mapping Phred quality score was slightly over 50, translating to a single base error probability estimation of 0.001%. Post-mapping yield of the run was just short of 895 Mbases and the general mapping error rate was estimated to be around 19%. The relational distribution between the read length, Phred quality score and read counts of the Seq. run 4 are depicted in Image 12.

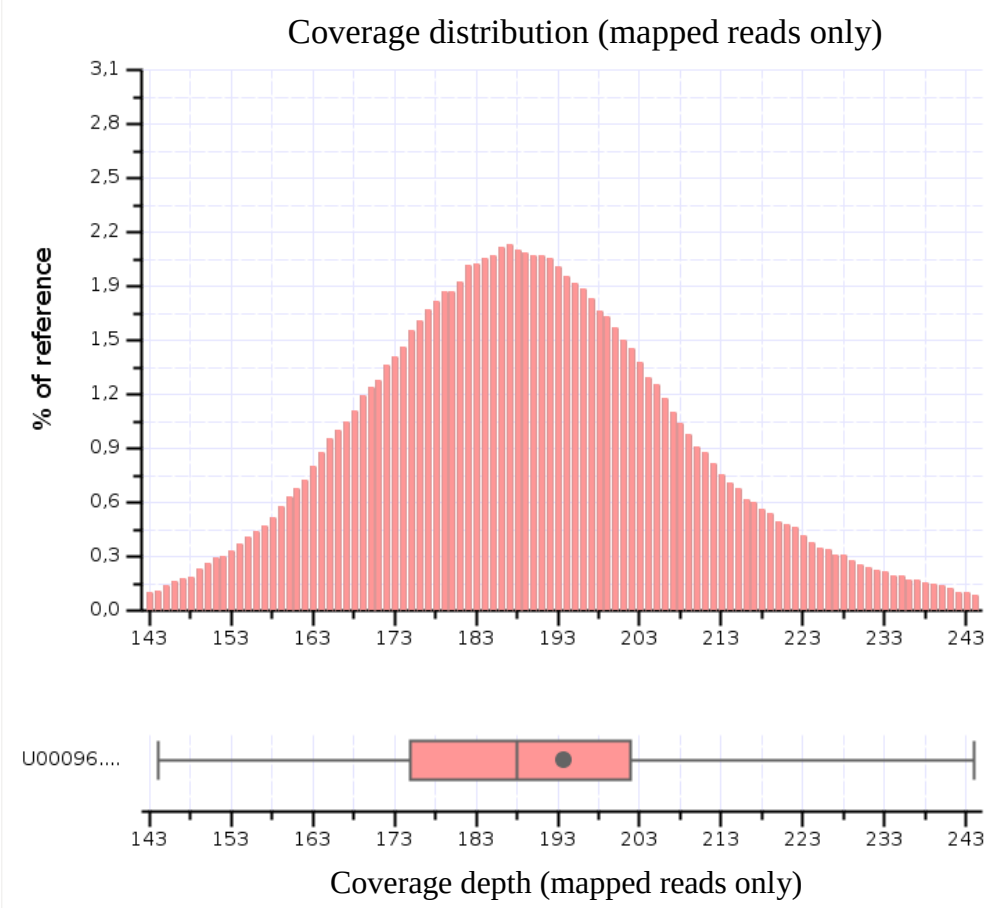

Image 11. The coverage depth of the post-alignment Seq. run 4 reads over the reference genome

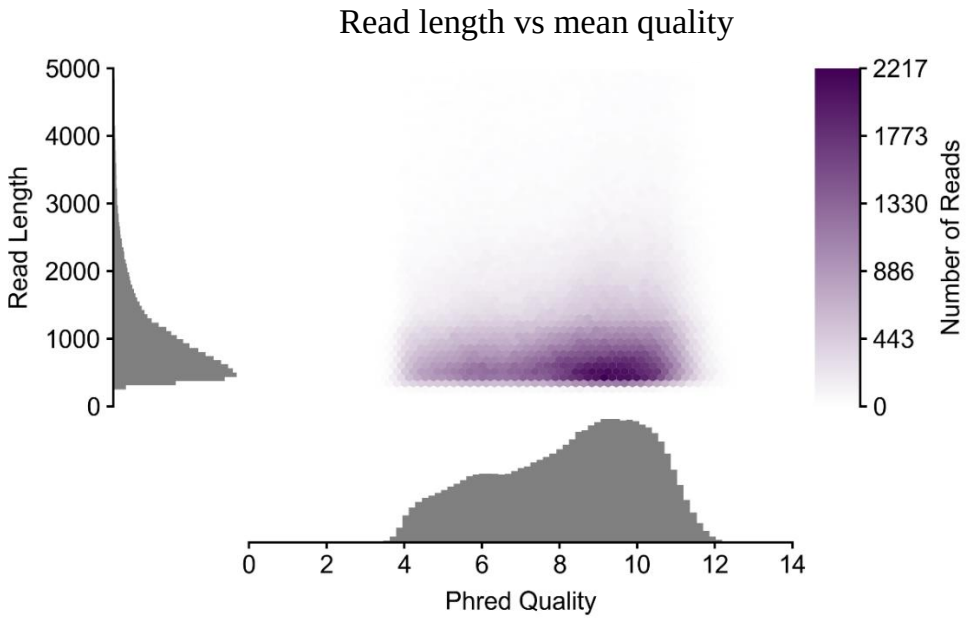

Image 12. The relational chart of the Seq. run 4 read characteristics.

4.4 General Data Analysis Results

It became soon apparent that directly adapting general sequencing analysis pipelines to the MinION data would not produce any meaningful results. The traditional alignment algorithms have not been optimized for aligning long nanopore reads with relatively high error rate. At first, the only aligner capable of producing any sort of meaningful alignments from the data was LAST. However, the alignment produced from our data with the LAST were far from good quality and combined with the poor yields from our early runs proved to have little actual research value. However, this issue was solved by the new alignment algorithm releases during the progression of the study. The introduction of the MinION nanopore support to the *BWA-MEM* algorithm allowed us to produce much more reliable and comprehensive alignments using our sequencing data. The *Graphmap* aligner released at an even later stage of our study also seemed to perform well. However, its release happened during such a late stage of our study that it was mostly implemented as an alternative option alongside *BWA* and not examined as an option for the main analysis pipeline.

We also observed there to be large difference in the median quality and mappability between the reads in Pass and Fail folders produced by the MinION. This observation led us into the decision of only including the reads placed in the Pass folder in our analyses. Including the failed reads into the result evaluation would only have muddled the dataset and produced very little meaningful information. Restricting the analysis to the passed reads generated much more reliable sequencing results that were still generally expansive enough for the requirements of the study.

We also considered going back to our old sequencing raw data with an intent of re-analyzing it with the help of improved basecalling and data analysis options. However, the backwards compatibility between old and new datasets proved to be quite limited due to the signal data differences produced by different sequencing setups. This made the prospect of re-analyzing old raw data for the most part either impossible or unproductive. We ultimately decided not to pursue this any further and focused our efforts on the production and analysis of new sequencing datasets.

Finally, the possibility of fully automating the general data analysis and filtering processes through master scripts combining different analysis steps together was also explored but later abandoned. While this approach would simplify the process of analysis and guarantee direct comparability between different sequencing datasets, it would not have been flexible enough for the requirements of this study. Over the course of the study we observed fluctuations in the read quality, constant changes to the MinION sequencing specifics and repeated updates to the analyzation tools. By flexibly adapting our sequencing and analysis pipelines in response to these changes we were able to better react to the rapidly evolving MinION sequencing method. Taking this into account, the benefits of an automated analysis pipeline in the context of this study would have been limited as best. We will reassess the benefit of analysis automation again once the MinION technology has matured more and we have reached a sequencing performance we are satisfied with.

5 DISCUSSION

Establishing operational capability with the Oxford Nanopore MinION sequencer turned out to be much more challenging than we initially anticipated. During the thesis project the practical challenges of adapting the MinION device into a new sequencing environment became apparent and their influence remained strong over the entire process. While the initial cost requirements for the device itself are quite low, the additional investments required for building the required complimentary research infrastructure are not trivial. Luckily, many of the required complementary tools such as a suitable sequencing computer may already be readily available for most researchers without a need for additional investments. This may not always be the case however, and especially for smaller research groups the associated supplementary purchases may well exceed the cost of the main MinION unit itself. Outside the associated monetary costs, a proper configuration and compatibility testing of the computing environment for the MinION usage requires some investment in the form of time. The process also requires some know-how, although the general setup of the system overall is fairly simple.

Another matter entirely is building the analysis pipeline for the sequencing data. As is the case with all NGS data, the analysis of the data places quite notable computing power requirements for the working environment. For most groups the most optimal solution would likely be either to utilize a dedicated computing cluster or to outsource the data analysis. While outsourcing is a reliable and easy way of performing the data analysis, it does incur additional costs tied to the sequencing. Finding a suitable service provider for data analysis of the still relatively uncommon MinION reads may also prove difficult. On the other hand, if the data analysis is done in-house and no suitable computing cluster is readily available the costs of procuring a heavy-duty computing environment could easily multiple the total cost of the sequencing environment.

The early results of the MinION test runs were not promising and the first experiments on the device did not produce any meaningful sequence data. No obvious explanation for these results was available either, making troubleshooting difficult. The device was

still officially unreleased at this point however, meaning it would have been premature to cast conclusive judgement upon the sequencer. Instead, the MinION pipeline development was planned based on the pace the improvements for the MinION became available.

While the largely inconsistent sequencing results during the early burn-in experiments initially looked peculiar, there are multiple feasible explanations for them. The probability of human error cannot be conclusively ruled out and it is a likely explanation for at least some of the fluctuation of the read yields. Moreover, the novelty and the constant updates to the technology were also behind many of the challenges. The original burn-in experiments of this study were among the first sequencing runs in the entire world done on the MinION device in a real research environment. At the time many of the intricacies of the system were also kept confidential and hidden from the users, including but not limited to the detailed structure of the sequencing chamber of the flow cell.

Worthy of special mention is the possibility that some of the early sequencing experiments were underperforming because of the membrane rupture inside the flow cells. This can happen either because of too aggressive or otherwise non-optimal pipetting technique during the priming of the flow cell and at the sample loading. The membrane rupture leads to a direct loss of sequencing capacity and yield of the entire run because the pores attached to a ruptured membrane are permanently lost. Although the pipetting steps required in the MinION flow cell sample handling are relatively simple, they require a certain amount of precision regarding the pipetting velocity. This kind of requirement is unusual in the common laboratory work, meaning it was not taken into consideration during the early experiments. This kind of consideration was also not common knowledge in the MinION userbase during the MAP phase. Thus it is quite likely that incorrect pipetting measures were employed at times, causing unintended damage to the flow cells.

Another possible cause for this kind of membrane destruction can be the introduction of air into the sequencing chamber. This can be caused by accidental insertion by pipetting or an error during the flow cell production. It was quite common for the early flow cells to have air bubbles inside their sequencing chambers when they were received by the end user. Many of the flow cells used during this thesis project also shared this characteristic, indicating they may have already lost much of their sequencing capacity even before they were used for sequencing. How much these factors ultimately contributed to the sequencing capability of the MinION were also not well-documented for a long time by the MAP. This explains why many of the experiments were performed with such non-optimal equipment.

The poor sequencing results and the continuous development of the technology led to long intervals between repeat attempts. It is important to note however, that for the majority of this time the pipeline development was not done full-time. The long downtime between various sequencing runs could also be used to study the technology and the characteristics of the produced reads and to prepare for the future experiments. Furthermore, during this time a functional bioinformatics environment for the data analysis was also established from scratch, including the procurement of necessary local computing hardware and an access to a suitable high-performance computing (HPC) environment.

In the end, the overall time between the first burn-in experiment and first real sequencing experiment took more than year. This is a considerably long time even for

establishing a new sequencing pipeline. Not until June 2015 the burn-in experiments were concluded. Only at this time, the reads passed the preliminary quality assessment performed during the basecalling. Although the yields at this point were still low, it was determined prudent to test the technology on real samples. Only from this point forward, the capabilities of the MinION device could also be truly assessed.

Those familiar with the traditional sequence alignment pipelines may wonder why no adapter or read trimming were utilized in any part of the thesis project. While this would undoubtedly have improved the overall alignments of the reads, the step was intentionally left outside the analyses discussed in this thesis. At the time the early experiment runs were performed, the optimal method for trimming the MinION reads was still a topic of uncertainty. Furthermore, whether the adapter sequences would be present in the final reads to the extent of justifying adapter trimming was uncertain. Since then, the MinION Nanopore community has reached an agreement on both accounts and the current recommendation is to precede the alignment with the adapter trimming step using a free program *porechop*[45] or other alternative. While this option will certainly be explored more closely in the future experiments with the MinION, it was excluded from this thesis.

The results of the first sequencing run performed using our own sample were very similar to those of the Burn-in 9. From this point onward a steady improvement to the overall performance of the system could be observed with each consequent experiment, excluding those implementing new protocol elements. The most likely explanation behind this sudden shift in performance is the compound effect of ONT updates to the technology and the optimizations to the sequencing pipeline implemented by us.

Moving on to the targeted sequencing, the initial results were again less than impressive. The first targeted sequencing experiments Targ. run 1 and 2 resulted again into a much lower read counts than the preceding sequencing run. However, the sequences that were produced passed the Metrichor filtering and appeared to be of

reasonably good quality. We speculated this to be an indication of the problem residing in the targeting steps and not in the sequencing process itself.

While it was certainly possible that there was a bigger problem with the xGen targeting protocol, our hypothesis was that the problem lay in the sample composition. Our reasoning was that the concentration of the ROI target in the entire human gDNA sample was so small, the targeting probes could not effectively locate and capture them with the reagent amounts used. Alternatively, the output yield from targeting protocol could simply be too minor to be effectively sequenced with the MinION protocol.

To test our theory we repeated the targeted sequencing protocol with the plasmid insert samples. The results from these experiments Targ. run 3 and 4 seem to at least partially confirm our hypothesis. Not only was the read yield much improved compared to the previous two targeting experiments, but the alignments matched the targeted regions splendidly. Based on these results, we concluded that the basic principle behind our targeting protocol was indeed functional. Unfortunately, implementing it for a human genome would be difficult. Especially problematic would be the clinical samples that are typically of lower DNA quality and purity than the ones we extracted ourselves for our experiments.

The unique characteristics of the MinION will establish the technology in its own place and purpose next to the conventional sequencing methods and other NGS solutions. The theoretically unrestricted read length allows the MinION to establish its own unique niche, especially in the research of the more challenging and repetitive regions of the genome. The low initial investment needed to establish the sequencing environment allows the technology to spread outside of the sequencing core facilities and opens interesting possibility in the educational field. The small spatial footprint combined with the relative simplicity and versatility of the library preparation protocol allow unprecedented flexibility in the sequencing environment. Combined, these characteristics pave the way for entirely new possibilities when it comes to sequencing both in the laboratory environment and the field studies.

Even with all the positive characteristics of the MinION, it is important to note the shortcomings of the technology. While the MinION sequencing has been evolving at an impressing rate the past few years, the technology currently remains inferior to its competitors in many ways. To date, the quality of the MinION sequencing data is orders of magnitude below that of the conventional sequencing methods such as Illumina sequencing by synthesis. The overall yield of a single sequencing run also remains rather unimpressive. Furthermore, the capability of high length read production is not unique to MinION with technologies such as Pacific BioSciences (PacBio) Single Molecule, Real-Time (SMRT) sequencing existing as a direct competitor.[46,47] It is also difficult to reliably predict how much longer the MinION sequencing can continue to improve and how the larger sequencing community will react to it. As it stands, every research project utilizing MinION will be operating with a certain level of uncertainty due to the state of the MinION and Nanopore sequencing as a whole.

While the MinION sequencer has been used to produce an entire human genome assembly both in-house and in the field, these kind of projects are still resource-heavy and expensive endeavours.[48, 49] Perhaps a more meaningful way of looking at the MinION sequencing at this point would be to consider it as a supplementary sequencing option instead of a direct competitor to the other technologies. Multiple research groups around the world have started genome assembly projects combining Illumina and Nanopore sequencing data to offset the downsides of both datasets. The leading principle of these projects is to match the accurate but short-length Illumina data against the lengthy but error-prone Nanopore reads. This way the errors of the Nanopore reads can be corrected through the more accurate Illumina reads, which in turn can be much more accurately assembled thanks to the much longer Nanopore reads. The results so far have led to some impressive genome assemblies and show a lot of promise in this field. One such example is the *Saccharomyces cerevisiae* assembly experiment by Goodwin *et al*. Using the hybrid assembly and error correction method combining MinION and MiSeq reads they were able to produce a single-contig assembly with over 99.99% identity. The same assembly constructed using the Illumina data alone was heavily fragmented with hundreds of contigs.[50]

The hybrid assembly is but one example of how the MinION sequencing could be implemented into existing research projects. Environmental observation efforts could use the portability and short turn-over time of the MinION to perform fast preliminary assessment of the local microbiota before moving to more focused sequencing with more accurate sequencing methods. A recent example of such project is the biodiversity assessment effort of the rainforest by Pomeranz *et al*.[51] The same basic principle can also be used for fast preliminary diagnosis confirmation or outbreak observation such as with the ebola project by Quick *et al.*[52]

The closer examination of the typical of MinION sequencing targets and the current industry standard Illumina alludes to the fact that the two technologies may not be in a direct competition with each other as much as one might think. The Nanopore sequencing would be the most beneficial for either solving the repetition counts and other large genomic variations or for producing highly intact transcriptome sequence. Both applications are something that Illumina and many other sequencing technologies struggle with due to their limited read length. Conversely, the limited base reliability of the MinION sequences hinders their use in solving SNPs and other minute variations. Research focusing on this kind of research questions will most likely continue using the well-established and reliable conventional sequencing pipelines in the future.

However, the currently existing situation of peaceful co-existence between MinION, Illumina and other NGS technologies may be a fleeting one. The distribution of different sequencers in the field of genomic research is all but evenly balanced; with the vast majority of all sequencing data is produced using various Illumina devices. This has already made it difficult for new sequencing technologies to gain enough foothold in the market to truly break through. The MinION will not even be able to solely rely on the novelty of long reads, as the PacBio system offers similar capabilities. While the closer inspection proves the vast differences between all the technologies, the superfluous similarities between them may often be just as important for possible future users in project planning stages.

The situation may also tip towards the other way. The nanopore sequencing itself avoids many of the limiting factors hard-coded to the functionality of other sequencers. If either MinION or some other future device will be able to continue advancing the technology further, the sequencing field may see drastic changes in the future. Just in the span of this thesis, the MinION has evolved from barely functional sequencer into an already established method for specific research situations. Notably, the theoretical limits of this evolution can still not be seen.

Many of the currently limiting factors of the MinION have been accurately identified and documented well. Many of these issues should be addressable to an extent through a combination of more advanced hardware and software, given enough time. Undoubtedly, the MinION will not be able to fully solve these issues, but some level of alleviation is to be expected. Whether this development is strong and fast enough for the MinION to impress those currently not convinced by the technology is another matter entirely.

If the quality of the nanopore reads would eventually improve enough to match that of Illumina and other short-read sequencers, the ramifications on the sequencing market would be massive. After all, longer reads are always preferable and more informative than shorter ones, given the two are identical in other aspects. Depending on the type of errors, the bigger read length may well become more valuable than the quality if high enough reliability threshold is achieved.

The MinION and nanopore sequencing have proven themselves as veritable sequencing options. The results of this thesis as well as those by many other researchers around the world have shown that MinION can produce reliable sequencing data. This opens the doors to not only nanopore-based DNA and RNA sequencing but wider range of nanopore sensing as a research method. Due to the way the nanopore signal data is produced, the basecalling can be expanded to infer more than just the nucleotide type. Currently there are tools available to observe DNA methylation from the non-basecalled

data, with the possibility of more sophisticated analysis options being developed in the future.[53]

The actions of ONT as a company going forward are also extremely important. Quite recently they released the PromethION sequencer, offering a much higher-throughput option for nanopore sequencing. New solutions like this, catering to the specific needs that the MinION is ill-fitted for are integral in expanding the availability and awareness of nanopore sequencing as a whole. The company has also been quite vocal in their plans of continuing to expand the technology through other new devices and additional protocols in the future.

Whether ONT really manages in securing a position in the highly contested sequencing field remains to be seen. While the nanopore sequencing itself could be considered disruptive technology with untapped potential to change an entire field of research, the MinION might easily remain unnoticed by the larger scientific community. Regardless, the experts of the field are sure to observe the progression of MinION closely.

6 CONCLUSIONS

The results obtained from this thesis research clearly confirm the potential of the Oxford Nanopore MinION sequencer as an alternative NGS option. It was clearly demonstrated to be capable of producing sequence that can be successfully aligned against a reference genome. The basic operations of the sequencing and analysis pipelines were also established and confirmed through multiple experiments.

The downsides of adopting a brand-new technology into research environment became clear during the project. The time the MinION was in its limited-availability MAP phase, both the sequencing and data analysis were performing considerably poorly. On the chemistry side, this could be attributed into two factors: the still not optimized library preparation kit and our unfamiliarity with the protocol. Although both have seen considerable improvement, others looking to establish their own sequencing pipeline for MinION should expect similar issues.

The data analysis for MinION reads proved to be considerably challenging as well. The length and unique error profile of the MinION reads caused many traditional analysis tools to underperform or outright malfunction. Luckily, this aspect of the pipeline has improved the most. Most of the key problems regarding the data analysis present at the beginning of this study have been solved by now. Nowadays, there are a plethora of available analysis tools for the alignment, polishing, trimming and analyzation of the MinION reads, all readily available for anyone. A lot of work still needs to be done on the subject, but the progress so far is extremely encouraging.

The cumulative effect of these factors led to considerable delays in the construction of the sequencing pipeline. This project is a prime example on how challenging it can be to estimate the effort needed in adopting a new technology. A similar project started today would likely be much more straightforward in progression than ours. The plethora of improvements on the MinION technology have addressed many of the challenges

depicted in this thesis. Still, it is unlikely that none of the challenges we have described would be shared by the future researches adapting the technology.

The overall trend of the MinION work becoming consistently easier and better established over the progress of the study is undeniable. The companion protocols and data analysis tools for the MinION data have become more readily available. Multiple sample datasets of experimental data is available for the tool development and the assessment of the technology. The library preparation process has been dramatically streamlined from the early days of MAP and can be completed faster and more consistently than before. The stability and reliability of the sequencing has improved tremendously due to the improvements to the technology and new bioinformatics tools. Best of all, these improvements can be clearly observed as in the sequencing results.

It seems logical to conclude that many of the challenges during this project can be attributed to the immaturity of the technology. While it could be said in hindsight that it would have been more time-efficient to wait until the official release of MinION before starting this project, there are unique benefits to the early adoption as well. Having been in close contact with the MinION over its development process has allowed for a much more intimate familiarity with the technology than otherwise possible. We could observe the effects that the protocol changes had on the sequencing process in real time and better estimate the future limits of the technology based on its development. This kind of experience provides an entirely new viewpoint and understanding of the entire technology and is often extremely difficult to obtain for more mature devices.

The next logical step for the research would be to expand towards the more specialized sequencing with the MinION. The experiments in this thesis project have irrefutably proven the capabilities of the MinION as a sequencer but also proven the necessity of target enrichment. The research of specific regions of larger genomes absolutely requires either a method for selective sequencing or notable investment through comprehensive high-depth sequencing of the entire human genome. The latter option is not realistically applicable for any tool hoping to see widespread research or clinical

use, leaving only the option of target enrichment. The preliminary experiments done during this thesis on this subject also show that this kind of target selection is possible, although challenging. Similar results have also been reported within the Nanopore community by other groups working with the MinION.

The natural progression for the study going forward is focusing more efforts on further target enrichment experimentation. While the initial experiments with the xGEN probe enrichment were not successful, the small-scale experiment performed on nebulin minigenes proved the validity of the method. Alternative probe designs or better optimization of the enrichment protocol may facilitate the full-scale implementation later. Another possibility is looking for alternative targeting methods easier to integrate into the MinION sequencing process. Additionally, the MinION technology can be branched out to include RNA sequencing and base modifications, such as methylation analysis. Protocols and analysis tools for both of these applications are already available, making integrating them into our research plan relatively simple.

ONT has also publicly acknowledged the importance of expanding the repertoire of the MinION. They have discussed targeted sequencing on multiple occasions and announced that targeting solutions are currently under internal development. [54, 55, 56] The furthest developed solution is currently designed to utilize the combination of Cas9 enzymes and RNA probes to bind the enzyme near the targeted region. This enzyme would then transfer the DNA to the sequencing pores, highly increasing the ratio of targeted sequence compared to the background.

ACKNOWLEDGEMENTS

I would like to extend my heartfelt thanks to the supervisors of my thesis project, Dr Katarina Pelin and Dr Kirsi Kiiski. Their help and guidance has been invaluable both for the completion of this thesis and for me personally. I will be eternally grateful I was given a chance to complete my master's thesis working on this interesting project. Having such a groundbreaking and forward-looking subject as my thesis project has truly been a blessing.

Besides my supervisors, I would like to thank the Folkhälsan Research Center for providing such an inspiring and pleasant working environment for my thesis project. My thanks go especially to the group leader of the Nemaline Myopathy Research Group at the Folkhälsan Research Center, Dr Carina Wallgren-Pettersson, and the rest of Nemaline Myopathy Research Group members Jenni Laitila, Vilma Lehtokari, Johanna Lehtonen, Lydia Sagath and Marilotta Turunen, as well as all the others I have had the pleasure of working with. Thanks to them I have been able to learn and experience more than I could ever have expected during this thesis project.

Additionally, I would like to give my thanks to the University of Helsinki Tumor Genomics research group of Lauri Aaltonen and the University of Helsinki Molecular Genetics of Immunological Diseases research group of Päivi Saavalainen. The experiences and knowledge these groups have been willing to share with us regarding the Nanopore technology has often been helpful beyond words. I would like to specifically thank Tiira Johansson and Päivi Saavalainen of the Saavalainen group and Saija Ahonen, Justyna Kolakowska and Kimmo Palin of the Aaltonen group for the help I've received from them.

Finally, I would like to thank my family and friends for being patient and supportive of my work and studies all this time.

REFERENCES

1. Watson, James D., and Francis HC Crick. "The structure of DNA." Cold Spring Harbor symposia on quantitative biology. Vol. 18. Cold Spring Harbor Laboratory Press, 1953.
2. Church, G., Deamer, D., Branton, D., Baldarelli, R., and Kasianowicz, J. (1998). Characterization of Individual Polymer Molecules Based on Monomer-Interface Interactions. US Patent No. 5795782, Brookline; Santa Cruz; Lexington; Natick; Darnestown.
3. Pelin, Katarina, et al. "Nebulin mutations in autosomal recessive nemaline myopathy: an update." Neuromuscular Disorders 12.7 (2002): 680-686.
4. Güttsches, Anne K., et al. "Two novel nebulin variants in an adult patient with congenital nemaline myopathy." Neuromuscular Disorders 25.5 (2015): 392-396.
5. Lehtokari, Vilma- Lotta, et al. "Mutation update: the spectra of nebulin variants and associated myopathies." Human mutation 35.12 (2014): 1418-1426.
6. Kiiski, Kirsi, et al. "A recurrent copy number variation of the NEB triplicate region: only revealed by the targeted nemaline myopathy CGH array." European Journal of Human Genetics 24.4 (2016): 574.
7. Kiiski, Kirsi. "Assessment of copy number variations in the nebulin gene and other nemaline myopathy-causing genes." (2015).
8. Donner, Kati, et al. "Complete genomic structure of the human nebulin gene and identification of alternatively spliced transcripts." European Journal of Human Genetics 12.9 (2004): 744.
9. Oxford Nanopore Technologies ONT Company history, Oxford Nanopore Technologies, https://nanoporetech.com/about-us/history, accessed on May 2018
10. University of California, Santa Cruz: New DNA sequencer uses nanopore concepts pioneered at UCSC, https://news.ucsc.edu/2012/02/nanopore-sequencer.html, accessed on May 2018
11. Kasianowicz, John J., et al. "Characterization of individual polynucleotide molecules using a membrane channel." Proceedings of the National Academy of Sciences 93.24 (1996): 13770-13773.
12. https://nanoporetech.com/how-it-works, accessed May 2018

13. de Lannoy, Carlos, Dick de Ridder, and Judith Risse. "The long reads ahead: de novo genome assembly using the MinION." F1000Research 6 (2017).
14. Loman Labs: Thar she blows! Ultra long read method for nanopore sequencing, published March 09 2017, accessed may 2018
15. Payne, Alex, et al. "Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files." bioRxiv (2018): 312256.
16. Simpson, Jared T., et al. "Detecting DNA cytosine methylation using nanopore sequencing." nature methods 14.4 (2017): 407.
17. Stoiber, M.H. et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv (2016).
18. Front line Genomics: PacBio and Oxford Nanopore Settle European Patent Dispute, http://www.frontlinegenomics.com/news/22688/pacbio-and-oxford-nanopore-settle-european-patent-dispute/
Published on May 10 2018, accessed on May 2018
19. Homolog.us; Business analysis - Oxford Nanopore, http://homolog.us/blogs/blog/2016/12/16/oxford-nanopore/, accessed on May 2018
20. Oxford Nanopore Technologies ONT, published information, accessed through ONT online community available at https://community.nanoporetech.com
21. Lahiri, Debomoy K., and John I. Nurnberger Jr. "A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies." Nucleic acids research 19.19 (1991): 5444.
22. Sagath, L., et al. "Functional studies of YBX3 variants associated with nemaline myopathy." *Neuromuscular Disorders* 26 (2016): S134-S135.
23. https://eu.idtdna.com/site/order/ngs
24. CSC – IT Center for Science, Finland: Taito-Shell HPC environment, https://research.csc.fi/taito-user-guide
25. CSC – IT Center for Science, Finland, https://research.csc.fi/home
26. Loman, Nicholas J., and Aaron R. Quinlan. "Poretools: a toolkit for analyzing nanopore sequence data." Bioinformatics 30.23 (2014): 3399-3401.
27. Poretools documentation, https://poretools.readthedocs.io/en/latest/, accessed on May 2018
28. Watson, Mick, et al. "poRe: an R package for the visualization and analysis of nanopore sequencing data." Bioinformatics 31.1 (2014): 114-115.

29. Babraham Bioinformatics, FastQC project main page, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, accesed on May 2018
30. Matthew Shirley, Fastqp: Fastqp project main page, https://github.com/mdshw5/fastqp, accessed on May 2018
31. Kiełbasa, Szymon M., et al. "Adaptive seeds tame genomic sequence comparison." Genome research 21.3 (2011): 487-493.
32. Last project main page http://last.cbrc.jp/, accessed on May 2018
33. BWA project main page, http://bio-bwa.sourceforge.net/, accessed on May 2018
34. Sović, Ivan, et al. "Fast and sensitive mapping of nanopore sequencing reads with GraphMap." Nature communications 7 (2016): 11307.
35. GraphMap project main page, https://github.com/isovic/graphmap, accessed on May 2018
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9
37. Samtools project main page, http://www.htslib.org/, accessed on May 2018
38. S. Andrews, Babraham Bioinformatics, BamQC project main page, https://github.com/s-andrews/BamQC, accessed on May 2018
39. James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)
40. Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).
41. Broad Institute, IGV project main page, http://software.broadinstitute.org/software/igv/home, accessed on May 2018
42. Katainen, Riku, et al. "BasePlayer: Versatile Analysis Software For Large-Scale Genomic Variant Discovery." bioRxiv (2017): 126482.
43. Aaltonen lab, University of Helsinki, BasePlayer main page, https://baseplayer.fi/, accessed on May 2018

44. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402
National Center for Biotechnology Information, Blast project main page https://blast.ncbi.nlm.nih.gov/Blast.cgi, accessed on May 2018
45. https://github.com/rrwick/Porechop, accessed on May 2018
46. Rhoads, Anthony, and Kin Fai Au. "PacBio sequencing and its applications." Genomics, proteomics & bioinformatics 13.5 (2015): 278-289.
47. https://www.pacb.com/smrt-science/smrt-sequencing/, accessed on April 2019
48. Clive G. Brown, Cliveome ONTHG1 data release, https://github.com/nanoporetech/ONT-HG1, published on 12 Dec 2016, accessed on May 2018
49. Jain, Miten, et al. "Nanopore sequencing and assembly of a human genome with ultra-long reads." Nature biotechnology 36.4 (2018): 338.
50. Goodwin, Sara, et al. "Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome." *Genome research* (2015).
51. Pomerantz, Aaron, et al. "Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building." GigaScience 7.4 (2018): giy033.
52. Quick, Joshua, et al. "Real-time, portable genome sequencing for Ebola surveillance." Nature 530.7589 (2016): 228.
53. Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nature methods, 14(4), 407.
54. https://nanoporetech.com/resource-centre/targeted-amplification-free-dna-sequencing-using-crisprcas9, accessed on April 2019
55. Gilpatrick, T. Cas9 targeted enrichment for nanopore profiling of methylation at known cancer drivers. Presentation. Available at: https://nanoporetech.com/resource-centre/cas9-targeted-enrichment-nanopore-profiling-methylation-known-cancer-drivers, accessed April 2019
56. Gabrieli, T. et al. Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH). Nucleic Acids Res. gky411 (2018).

ADDITIONAL DATA

Additional data 1: Reference sequences

```
>DNA_CS

GCCATCAGATTGTGTTTGTTAGTCGCTTTTTTTTTTTTGGAATTTTTTTTTTTGGAATTTTTTTTTTGCGCTA
ACAACCTCCTGCCGTTTTGCCCGTGCATATCGGTCACGAACAAATCTGATTACTAAACACAGTAGCCTGGA
TTTGTTCTATCAGTAATCGACCTTATTCCTAATTAAATAGAGCAAATCCCCTTATTGGGGGTAAGACATGA
AGATGCCAGAAAAACATGACCTGTTGGCCGCCATTCTCGCGGCAAAGGAACAAGGCATCGGGGCAATCCTT
GCGTTTGCAATGGCGTACCTTCGCGGCAGATATAATGGCGGTGCGTTTACAAAAACAGTAATCGACGCAAC
GATGTGCGCCATTATCGCCTAGTTCATTCGTGACCTTCTCGACTTCGCCGGACTAAGTAGCAATCTCGCTT
ATATAACGAGCGTGTTTATCGGCTACATCGGTACTGACTCGATTGGTTCGCTTATCAAACGCTTCGCTGCT
AAAAAAGCCGGAGTAGAAGATGGTAGAAATCAATAATCAACGTAAGGCGTTCCTCGATATGCTGGCGTGGT
CGGAGGGAACTGATAACGGACGTCAGAAAACCAGAAATCATGGTTATGACGTCATTGTAGGCGGAGAGCTA
TTTACTGATTACTCCGATCACCCTCGCAAACTTGTCACGCTAAACCCAAAACTCAAATCAACAGGCGCCGG
ACGCTACCAGCTTCTTTCCCGTTGGTGGGATGCCTACCGCAAGCAGCTTGGCCTGAAAGACTTCTCTCCGA
AAAGTCAGGACGCTGTGGCATTGCAGCAGATTAAGGAGCGTGGCGCTTTACCTATGATTGATCGTGGTGAT
ATCCGTCAGGCAATCGACCGTTGCAGCAATATCTGGGCTTCACTGCCGGGCGCTGGTTATGGTCAGTTCGA
GCATAAGGCTGACAGCCTGATTGCAAAATTCAAAGAAGCGGGCGGAACGGTCAGAGAGATTGATGTATGAG
CAGAGTCACCGCGATTATCTCCGCTCTGGTTATCTGCATCATCGTCTGCCTGTCATGGGCTGTTAATCATT
ACCGTGATAACGCCATTACCTACAAAGCCCAGCGCGACAAAAATGCCAGAGAACTGAAGCTGGCGAACGCG
GCAATTACTGACATGCAGATGCGTCAGCGTGATGTTGCTGCGCTCGATGCAAAATACACGAAGGAGTTAGC
TGATGCTAAAGCTGAAAATGATGCTCTGCGTGATGATGTTGCCGCTGGTCGTCGTCGGTTGCACATCAAAG
CAGTCTGTCAGTCAGTGCGTGAAGCCACCACCGCCTCCGGCGTGGATAATGCAGCCTCCCCCCGACTGGCA
GACACCGCTGAACGGGATTATTTCACCCTCAGAGAGAGGCTGATCACTATGCAAAAACAACTGGAAGGAAC
CCAGAAGTATATTAATGAGCAGTGCAGATAGAGTTGCCCATATCGATGGGCAACTCATGCAATTATTGTGA
GCAATACACACGCGCTTCCAGCGGAGTATAAATGCCTAAAGTAATAAAACCGAGCAATCCATTTACGAATG
TTTGCTGGGTTTCTGTTTTAACAACATTTTCTGCGCCGCCACAAATTTTGGCTGCATCGACAGTTTTCTTC
TGCCCAATTCCAGAAACGAAGAAATGATGGGTGATGGTTTCCTTTGGTGCTACTGCTGCCGGTTTGTTTTG
AACAGTAAACGTCTGTTGAGCACATCCTGTAATAAGCAGGGCCAGCGCAGTAGCGAGTAGCATTTTTTTCA
TGGTGTTATTCCCGATGCTTTTTGAAGTTCGCAGAATCGTATGTGTAGAAAATTAAACAAACCCTAAACAA
TGAGTTGAAATTTCATATTGTTAATATTTATTAATGTATGTCAGGTGCGATGAATCGTCATTGTATTCCCG
GATTAACTATGTCCACAGCCCTGACGGGGAACTTCTCTGCGGGAGTGTCCGGGAATAATTAAAACGATGCA
CACAGGGTTTAGCGCGTACACGTATTGCATTATGCCAACGCCCCGGTGCTGACACGGAAGAAACCGGACGT
TATGATTTAGCGTGGAAAGATTTGTGTAGTGTTCTGAATGCTCTCAGTAAATAGTAATGAATTATCAAAGG
TATAGTAATATCTTTTATGTTCATGGATATTTGTAACCCATCGGAAAACTCCTGCTTTAGCAAGATTTTCC
CTGTATTGCTGAAATGTGATTTCTCTTGATTTCAACCTATCATAGGACGTTTCTATAAGATGCGTGTTTCT
TGAGAATTTAACATTTACAACCTTTTTAAGTCCTTTTATTAACACGGTGTTATCGTTTTCTAACACGATGT
GAATATTATCTGTGGCTAGATAGTAAATATAATGTGAGACGTTGTGACGTTTTAGTTCAGAATAAAACAAT
TCACAGTCTAAATCTTTTCGCACTTGATCGAATATTTCTTTAAAAATGGCAACCTGAGCCATTGGTAAAAC
CTTCCATGTGATACGAGGGCGCGTAGTTTGCATTATCGTTTTTATCGTTTCAATCTGGTCTGACCTCCTTG
TGTTTTGTTGATGATTTATGTCAAATATTAGGAATGTTTTCACTTAATAGTATTGGTTGCGTAACAAAGTG
CGGTCCTGCTGGCATTCTGGAGGGAAATACAACCGACAGATGTATGTAAGGCCAACGTGCTCAAATCTTCA
TACAGAAAGATTTGAAGTAATATTTTAACCGCTAGATGAAGAGCAAGCGCATGGAGCGACAAAATGAATAA
AGAACAATCTGCTGATGATCCCTCCGTGGATCTGATTCGTGTAAAAAATATGCTTAATAGCACCATTTCTA
TGAGTTACCCTGATGTTGTAATTGCATGTATAGAACATAAGGTGTCTCTGGAAGCATTCAGAGCAATTGAG
GCAGCGTTGGTGAAGCACGATAATAATATGAAGGATTATTCCCTGGTGGTTGACTGATCACCATAACTGCT
AATCATTCAAACTATTTAGTCTGTGACAGAGCCAACACGCAGTCTGTCACTGTCAGGAAAGTGGTAAAACT
GCAACTCAATTACTGCAATGCCCTCGTAATTAAGTGAATTTACAATATCGTCCTGTTCGGAGGGAAGAACG
CGGGATGTTCATTCTTCATCACTTTTAATTGATGTATATGCTCTCTTTTCTGACGTTAGTCTCCGACGGCA
GGCTTCAATGACCCAGGCTGAGAAATTCCCGGACCCTTTTTGCTCAAGAGCGATGTTAATTTGTTCAATCA
TTTGGTTAGGAAAGCGGATGTTGCGGGTTGTTGTTCTGCGGGTTCTGTTCTTCGTTGACATGAGGTTGCCC
CGTATTCAGTGTCGCTGATTTGTATTGTCTGAAGTTGTTTTTACGTTAAGTTGATGCAGATCAATTAATAC
GATACCTGCGTCATAATTGATTATTTGACGTGGTTTGATGGCCTCCACGCACGTTGTGATATGTAGATGAT
AATCATTATCACTTTACGGGTCCTTTCCGGTGAAAAAAAAGGTACCAAAAAAAACATCGTCGTGAGTAGTG
```

Data 1. The DNA-CS reference sequence.

>pGEX-4T1

ACGTTATCGACTGCACGGTGCACCAATGCTTCTGGCGTCAGGCAGCCATCGGAAGCTGTGGTATGGCTGTGCAGGTCGTAAATCACTG
CATAATTCGTGTCGCTCAAGGCGCACTCCCGTTCTGGATAATGTTTTTTGCGCCGACATCATAACGGTTCTGGCAAATATTCTGAAAT
GAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGTATTCATGTCCC
CTATACTAGGTTATTGGAAAATTAAGGGCCTTGTGCAACCCACTCGACTTCTTTTGGAATATCTTGAAGAAAAATATGAAGAGCATTT
GTATGAGCGCGATGAAGGTGATAAATGGCGAAACAAAAAGTTTGAATTGGGTTTGGAGTTTCCCAATCTTCCTTATTATATTGATGGT
GATGTTAAATTAACACAGTCTATGGCCATCATACGTTATATAGCTGACAAGCACAACATGTTGGGTGGTTGTCCAAAAGAGCGTGCAG
AGATTTCAATGCTTGAAGGAGCGGTTTTGGATATTAGATACGGTGTTTCGAGAATTGCATATAGTAAAGACTTTGAAACTCTCAAAGT
TGATTTTCTTAGCAAGCTACCTGAAATGCTGAAAATGTTCGAAGATCGTTTATGTCATAAAACATATTTAAATGGTGATCATGTAACC
CATCCTGACTTCATGTTGTATGACGCTCTTGATGTTGTTTTATACATGGACCCAATGTGCCTGGATGCGTTCCCAAAATTAGTTTGTT
TTAAAAAACGTATTGAAGCTATCCCACAAATTGATAAGTACTTGAAATCCAGCAAGTATATAGCATGGCCTTTGCAGGGCTGGCAAGC
CACGTTTGGTGGTGGCGACCATCCTCCAAAATCGGATCTGGTTCCGCGTGGATCCCCGGAATTCCCGGGTCGACTCGAGCGGCCGCAT
CGTGACTGACTGACGATCTGCCTCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTT
GTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTC
ACGTAGCGATAGCGGAGTGTATAATTCTTGAAGACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATG
GTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATC
CGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTAT
TCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCA
CGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTT
TTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGA
CTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGT
GATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAA
CTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAAC
GTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGA
CCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAG
CACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGAT
CGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCAT
TTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGT
CAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCT
ACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACT
GTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAG
TGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAAC
GGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACG
CTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACG
CCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATG
GAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGAT
TCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGG
AAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATAAATTCCGACACCATCGAATGGTG
CAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAATGTGAAACCAGTAACGTTATACGATGTC
GCAGAGTATGCCGGTGTCTCTTATCAGACCGTTTCCCGCGTGGTGAACCAGGCCAGCCACGTTTCTGCGAAAACGCGGGAAAAAGTGG
AAGCGGCGATGGCGGAGCTGAATTACATTCCCAACCGCGTGGCACAACAACTGGCGGGCAAACAGTCGTTGCTGATTGGCGTTGCCAC
CTCCAGTCTGGCCCTGCACGCGCCGTCGCAAATTGTCGCGGCGATTAAATCTCGCGCCGATCAACTGGGTGCCAGCGTGGTGGTGTCG
ATGGTAGAACGAAGCGGCGTCGAAGCCTGTAAAGCGGCGGTGCACAATCTTCTCGCGCAACGCGTCAGTGGGCTGATCATTAACTATC
CGCTGGATGACCAGGATGCCATTGCTGTGGAAGCTGCCTGCACTAATGTTCCGGCGTTATTTCTTGATGTCTCTGACCAGACACCCAT
CAACAGTATTATTTTCTCCCATGAAGACGGTACGCGACTGGGCGTGGAGCATCTGGTCGCATTGGGTCACCAGCAAATCGCGCTGTTA
GCGGGCCCATTAAGTTCTGTCTCGGCGCGTCTGCGTCTGGCTGGCTGGCATAAATATCTCACTCGCAATCAAATTCAGCCGATAGCGG
AACGGGAAGGCGACTGGAGTGCCATGTCCGGTTTTCAACAAACCATGCAAATGCTGAATGAGGGCATCGTTCCCACTGCGATGCTGGT
TGCCAACGATCAGATGGCGCTGGGCGCAATGCGCGCCATTACCGAGTCCGGGCTGCGCGTTGGTGCGGATATCTCGGTAGTGGGATAC
GACGATACCGAAGACAGCTCATGTTATATCCCGCCGTTAACCACCATCAAACAGGATTTTCGCCTGCTGGGGCAAACCAGCGTGGACC
GCTTGCTGCAACTCTCTCAGGGCCAGGCGGTGAAGGGCAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACCCTGGCGCC
CAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGA
GCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATT
GTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAA
AACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCC
CTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTTTGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTG
CGATCTTCCTGAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATCTACACCAACGTAACCTAT
CCCATTACGGTCAATCCGCCGTTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTAC
AGGAAGGCCAGACGCGAATTATTTTTGATGGCGTTGGAATT

Data 2. The reference sequence of pGEX-4T1 plasmid.

>NEBex53-57

TATAAATACAAACAAGGCTACCGAAAGCAACTTGGCCACCATGTTGGATTCCGGAGTCTGCAAGA
TGACCCAAAACTTGTGTTGTCCATGAATGTAGCCAAAATGCAGAGTGAAAGAGAATACAAGAAGG
ACTTTGAGAAGTGGAAAACTAAGTTCTCCAGCCCAGTGGACATGTTGGGAGTGGTACTGGCCAAG
AAGTGTCAGGAGTTGGTTAGTGACGTGGACTACAAGAACTACCTGCATCAGTGGACATGTCTGCCT
GATCAGAACGATGTTGTGCAAGCTAAGAAAGTTTATGAACTGCAAAGTGAGAATCTATATAAATC
TGACCTTGAGTGGCTGAGAGGCATAGGATGGAGTCCCTTGGGTTCTTTAGAGGCAGAAAAGAACA
AGCGGGCTTCGGAAATCATCAGTGAGAAGAAATATCGTCAGCCTCCAGACAGAAACAAGTTCACC
AGCATTCCTGATGCCATGGATATAGTTCTGGCAAAGACAAATGCCAAAAATAGGAGTGATAGACT
TTATAGAGAAGCTTGGGACAAAGACAAGACTCAGATCCACATCATGCCTGATACACCTGACATTG
TTCTGGCTAAAGCAAACTTAATCAACACAAGTGATAAACTCTACCGAATGGGTTATGAGGAGCTG
AAGAGAAAAGGTTACGATCTTCCTGTTGATGCCATACCAATCAAAGCAGCAAAAGCCTCCCGGGA
AATTGCCAGTGAATACAAGTACAAGGAAGGCTTTCGCAAGCAGCTCGGCCACCACATTGGTGCCC
GGAACATTGAAGATGACCCCAAGATGATGTGGTCCATGCATGTGGCCAAGATCCAGAGTGACAGG
GAGTACAAGAAGGACTTTGAGAAGTGGAAGACCAAGTTCAGCAGCCCAGTGGACATGCTGGGGGT
GGTGTTGGCCAAGAAGTGCCAGACCTTAGTCAGCGACGTGGACTACAAGAACTACCTGCACCAGT
GGACATGCCTGCCCGACCAGAGCGATGTCATCCATGCTCGGCAGGCCTATGACCTCCAGAGCGAT

Data 3. The reference sequence of NEB ex53-57 insert.

>NEBex77-81

TACAAGTACAAGGAAGGCTACCGCAAACAGCTTGGCCACCATATTGGGGCCCGGAACATTAAGG
ATGACCCGAAGATGATGTGGTCCATCCATGTGGCCAAGATCCAGAGTGACAGGGAGTACAAGAA
GGAGTTTGAGAAGTGGAAGACCAAGTTCAGCAGCCCAGTGGACATGCTGGGGGTGGTGCTGGCC
AAGAAGTGTCAGATCCTTGTAAGCGACATAGACTACAAGCATCCCCTGCATGAATGGACCTGCCT
GCCTGATCAGAATGACGTCATTCAGGCTCGGAAGGCCTATGACCTGCAGAGTGATGCTATTTACA
AATCTGATCTTGAGTGGCTGAGAGGCATAGGATGGGTTCCCATTGGCTCTGTAGAGGTCGAGAAA
GTGAAGAGAGCTGGAGAAATCCTGAGTGACAGGAAGTATCGCCAGCCTGCAGACCAGCTCAAAT
TCACATGCATTACCGACACTCCGGAAATTGTCCTAGCAAAGAATAATGCCCTGACAATGAGCAAG
CATTTATACACAGAAGCTTGGGATGCTGACAAAACCTCCATCCACGTGATGCCAGACACCCCAGA
TATCCTGCTGGCCAAGAGTAATTCTGCCAATATCAGCCAAAAACTTTACACCAAGGGATGGGATG
AATCAAAGATGAAGGACTATGATCTGAGAGCAGATGCTATTTCCATCAAAAGTGCCAAGGCCTCC
AGGGACATCGCCAGTGACTACAAATACAAGGAAGCCTATGAGAAACAGAAAGGCCACCACATTG
GAGCCCAGAGCATTGAAGATGATCCCAAGATTATGTGTGCCATACATGCAGGAAAAATTCAAAGT
GAAAGGGAGTACAAGAAGGAATTCCAAAAGTGGAAAACCAAGTTCTCTAGCCCAGTGGACATGT
TAAGCATCTTGCTGGCCAAGAAATGTCAGACTTTGGTCACTGACATTGATTATCGCAATTACCTGC
ATGAATGGACATGCATGCCGGATCAAAACGACATTATCCAAGCAAAAAAGGCCTATGACCTGCA
GAGTGAT

Data 4. The reference sequence of NEB ex77-81 insert.

>NEBex119-125

CTTAAATACAAAGAGACATATGAGAAGCAGAAAGGTCACTACCTGGCTGGAAAAGTGA
TCGGTGAATTCCCTGGTGTGGTTCACTGTCTGGATTTCCAAAAGATGAGGAGTGCGTTG
AACTACAGAAAACATTATGAGGATACCAAAGCAAATGTTCATATCCCCAATGACATGAT
GAATCACGTGCTGGCTAAAAGGTGCCAGTACATCCTCAGTGACCTGGAGTATCGACACT
ATTTCCACCAGTGGACGTCTCTTCTGGAAGAACCCAATGTTATACGCGTCCGAAACGCC
CAGGAGATCTTGAGTGATAATGTGTATAAAGATGACCTGAATTGGTTGAAAGGCATTGG
TTGCTACGTTTGGGATACACCCCAAATCCTCCATGCCAAGAAATCATACGACCTTCAGA
GTCAGCTACAATATACAGCAGCAGGTAAAGAAAATCTACAAAACTATAATCTGGTCACA
GACACGCCCCTCTATGTGACTGCTGTTCAGAGTGGCATTAATGCCAGTGAGGTAAAATA
TAAAGAAAATTATCATCAGATTAAGGACAAATACACAACAGTTCTAGAAACAGTGGATT
ATGACAGAACCAGAAACCTGAAGAATCTTTACAGCAGTAACCTGTACAAGGAGGCCTG
GGATAGAGTGAAAGCCACCAGCTACATCCTGCCTTCCAGCACCTTGTCCCTGACACACG
CCAAGAACCAGAAGCATCTGGCCAGCCATATCAAATATCGGGAAGAATATGAAAAGTT
CAAAGCTCTTTATACGTTACCAAGAAGTGTTGACGATGATCCGAACACAGCACGGTGCC
TCCGAGTTGGCAAGCTTAACATCGAT

Data 5. The reference sequence of NEB ex119-125 insert.

>NEBex146-153

ATCCTTTATAAATTGGAATACAACAAGGCCAAACCCAGAGGCTACACCACAATCCACGA
CACACCCATGTTGCTGCATGTCCGCAAGGTTAAAGATGAAGTCAGTGATCTGAAATACA
AAGAAGTATACCAAAGAAATAAATCCAACTGCACCATTGAGCCAGATGCTGTTCATATC
AAAGCAGCCAAGGACGCCTACAAAGTCAACACCAATCTGGACTATAAGAAACAGTACG
AAGCCAACAAAGCCCACTGGAAGTGGACTCCTGACCGACCGGACTTCCTCCAGGCTGCC
AAGTCATCCCTGCAGCAAAGCGATTTTGAATATAAGCTGGACCGGGAGTTCCTCAAGGG
TTGCAAGCTTTCTGTCACTGATGACAAAAACACGGTGCTCGCCCTCAGGAATACTTTAAT
AGAAAGTGATCTGAAATACAAAGAGAAACATGTCAAGGAAAGAGGAACCTGCCATGCC
GTACCTGACACGCCTCAGATCCTGCTGGCGAAGACTGTCAGCAACCTGGTGTCTGAGAA
CAAGTACAAGGACCATGTCAAGAAGCACTTGGCACAGGGCTCATACACAACACTACCAG
AGACCCGGGACACTGTTCACGTCAAGGAAGTGACCAAGCATGTCAGTGATACAAATTAC
AAAAAGAAGTTTGTCAAGGAGAAAGGAAAATCCAACTACTCCATCATGCTGGAGCCACC
AGAGGTGAAACATGCTATGGAAGTGGCCAAGAAGCAAAGTGATGTCGCTTACAGAAAA
GATGCCAAAGAGAACCTGCATTACACCACAGTGGCTGATCGACCAGACATCAAGAAGGC
CACACAGGCAGCCAAACAGGCCAGTGAG

Data 6. The reference sequence of NEB ex146-153 insert.

Additional data 2: Scripts and console commands

```
bwa index ref.fasta

bwa mem -x ont2d -M Reference.fasta SequencingReads.fastq
        > Alignments.sam
```

Data 3. Terminal commands for LAST alignment production.

```
parallel -fasta "lastal (-s 2 -T 0 -Q 0
        -a 1) ReferenceIndexName"
        < SequencingReads.fasta > Alignments.txt

maf-convert.py sam Alignments.txt > Alignments.sam
```

Data 4. Terminal commands for BWA alignment of MinION sequence data.

```
graphmap align -r Reference.fasta -d SequencingReads.fasta
        -o Alignments.sam
```

Data 5. Terminal command for the alignment of MinION data using graphmap.

```
samtools view -bS -T ReferenceSeq.fasta
        -o myalns.bam myalns.sam

samtools sort myalns.bam -o myalns_sorted.bam

samtools index myalns_sorted.bam
```

Data 6. Terminal commands for the processing of the sam alignment files.